

This International Student Edition is for use outside of the U.S.

2024 RELEASE

ESSENTIALS OF MARKETING ANALYTICS

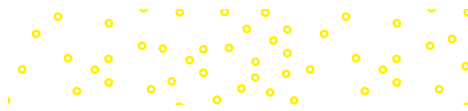
JOSEPH F.
HAIR, JR.

DANA E.
HARRISON

HAYA
AJJAN

Mc
Graw
Hill





Essentials of Marketing Analytics

2024 Release

Joseph F. Hair, Jr.

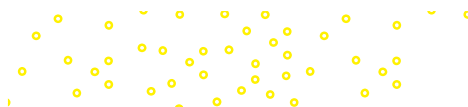
University of South Alabama

Dana E. Harrison

East Tennessee State University

Haya Ajjan

Elon University





ESSENTIALS OF MARKETING ANALYTICS

Published by McGraw Hill LLC, 1325 Avenue of the Americas, New York, NY 10019. Copyright ©2024 by McGraw Hill LLC. All rights reserved. Printed in the United States of America. No part of this publication may be reproduced or distributed in any form or by any means, or stored in a database or retrieval system, without the prior written consent of McGraw Hill LLC, including, but not limited to, in any network or other electronic storage or transmission, or broadcast for distance learning.

Some ancillaries, including electronic and print components, may not be available to customers outside the United States.

This book is printed on acid-free paper.

1 2 3 4 5 6 7 8 9 LKV 29 28 27 26 25 24

ISBN 978-1-266-93162-8

MHID 1-266-93162-7

Cover Image: *Ico Maker/Shutterstock*

All credits appearing on page or at the end of the book are considered to be an extension of the copyright page.

The Internet addresses listed in the text were accurate at the time of publication. The inclusion of a website does not indicate an endorsement by the authors or McGraw Hill LLC, and McGraw Hill LLC does not guarantee the accuracy of the information presented at these sites.

1

Introduction to Marketing Analytics

LEARNING OBJECTIVES

- 1.1** Define marketing analytics.
- 1.2** Discuss how to identify the right business problem.
- 1.3** Identify and compare different data sources.
- 1.4** Describe different data types.
- 1.5** Explain the difference between predictors and target variables.
- 1.6** Differentiate between supervised and unsupervised modeling.
- 1.7** Investigate the 7-step marketing analytics process.
- 1.8** Explain the value of learning marketing analytics.



alphaspirit/Shutterstock

1.1 Introduction to Marketing Analytics

A primary responsibility of marketing is to properly manage the wants and needs of customers. This can be accomplished through strategic decisions about products, pricing, distribution, and communications that are based on insights from marketing analytics. This chapter will introduce you to the exciting possibilities of marketing analytics, which companies are increasingly using to satisfy customers and maintain a competitive advantage.

Have you ever wondered how Hotels.com, Spotify, or Stitch Fix obtain and provide the information customers want so fast? As examples, consider the three situations below:

- **How does Expedia, Orbitz, or Hotels.com determine the price to quote when you are shopping for a hotel room?** Prices of hotel rooms are frequently updated based on demand, seasonality, day of the week, time of the day, and even the type of technology being used to find accommodations. For instance, Orbitz Worldwide Inc. knows that Mac computer users spend as much as 30 percent more a night on hotels, so Orbitz shows its Mac customers different travel options, and sometimes even more expensive rooms than Windows users.¹
- **How does Spotify know what songs to suggest for you?** From user-generated playlists, listener preferences, and advanced data analytics, Spotify, an audio streaming platform, can build collections of music their listeners enjoy and help users find their new favorite music.²
- **How does Stitch Fix achieve the highest-ever rate of purchased items per “Fix” for its female customers?** Stitch Fix started in 2011 and in 2022, it generated \$2.1 billion in sales. Their stylists work closely with the analytics algorithm suggestions, and then match results with the customer’s style. Over time, the analytics algorithm learns and continuously becomes more accurate when making clothing suggestions, stocking decisions, packing at the warehouse, and shipping.³
- **How does ChatGPT write an essay about marketing analytics in a few seconds?** ChatGPT, the popular chat platform from OpenAI reached 100 million active users in just two months after launch. It took Instagram 30 months to reach the same number of users. The AI chatbot can hold conversational text exchanges with users by using supervised and reinforcement machine learning methodology to make these exchanges feel as if you were chatting with a real person.

In the rest of this chapter, we describe and explain an analytics framework, the relevant marketing analytics concepts, and industry best practices. Building on this foundation, you will continue to work through practical exercises and develop the mindset of a marketing analyst.

Marketing Analytics Defined

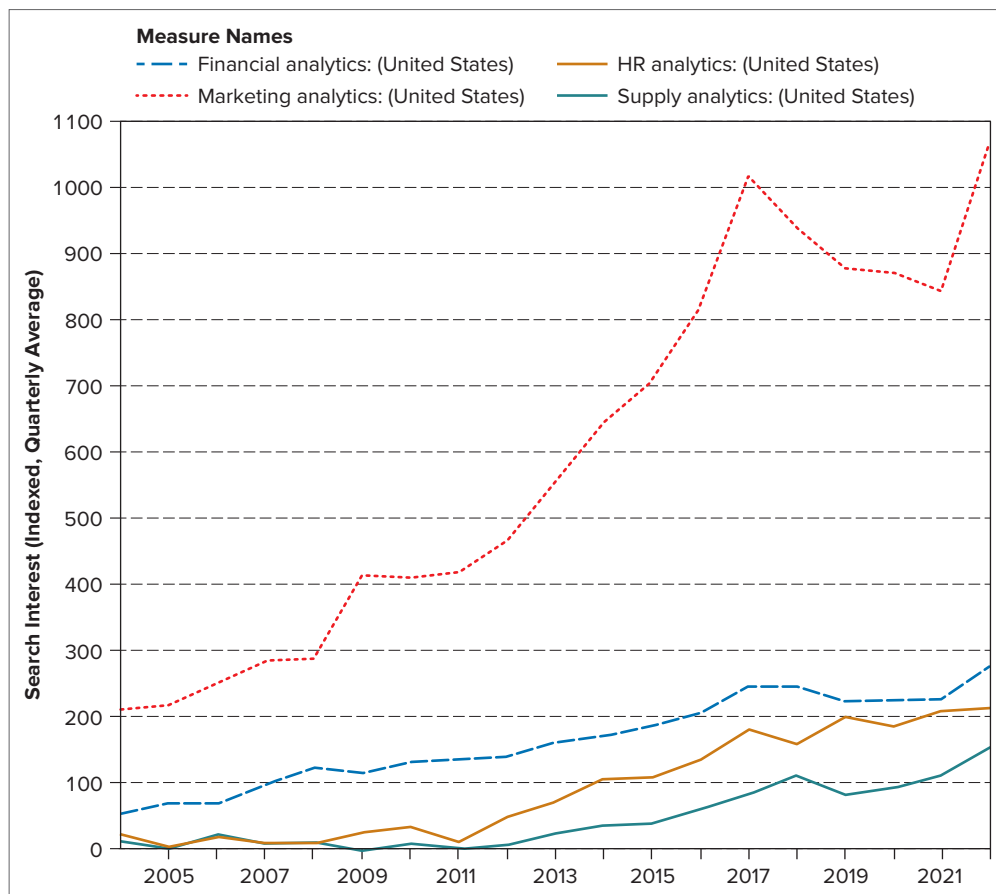
Marketing analytics uses data, statistics, mathematics, and technology to solve marketing business problems. It involves modeling and software to drive marketing decision making. Not long ago, marketing analytics was a highly specialized field for individuals who had in-depth knowledge of mathematical modeling, computer programming, and specialized software packages. Today, however, the availability of large amounts of data, improvements in analytics techniques, substantial increases in computer processing power, and affordability have made marketing analytics more practical and available to a much larger audience. To survive, companies increasingly need to differentiate products and services, optimize processes, and understand the drivers for business performance, and marketing analytics can help them to do that.

Marketing analytics is one of the fastest growing fields of analytics applications. This growth can be attributed to the increase in user-generated data from social media (e.g., Instagram, Facebook, Twitter), mobile applications (e.g., weather, text, maps), and multiple search and

shopping channels now accessible by customers (e.g., phone, in-store, online). Marketers can use insights from analytics to increase company performance through various marketing capabilities such as pricing, product development, channel management, marketing communications, and selling. Restaurants are even beginning to apply marketing analytics to optimize the selection of new locations. For example, the restaurant chain Roy Rogers Franchise Co. uses advanced analytics to expand into new markets, determine their next site locations, and forecast sales.⁴ Their machine learning platform integrates internal and external data to ensure restaurant locations match the needs and wants of the geographical area. Internal data such as the current location of stores, sales, and competitor locations are integrated with external data such as demographics, traffic near the store, and social media activity (e.g., geo-tagged posts) to gain a more holistic view of the site potential.

Marketing analytics is increasingly being applied in numerous industries and functional departments, and the impact and benefits are evident. Exhibit 1-1 compares the interest in marketing analytics to analytics use in other business functions. Results are measured based on the search volume for the word “Marketing Analytics” using Google Trends from 2004 to 2022. The search for marketing analytics has been consistently higher than other fields, with financial analytics, HR analytics, and supply chain analytics being much lower.

Exhibit 1-1 Google Search Trends for the Terms Marketing Analytics, Supply Chain Analytics, Financial Analytics, and HR Analytics (2004–2022)



Source: Google Trends.

A large amount of marketing data exists, which explains the interest in learning more about the area. A lack of marketing analytics skills, however, has left many companies

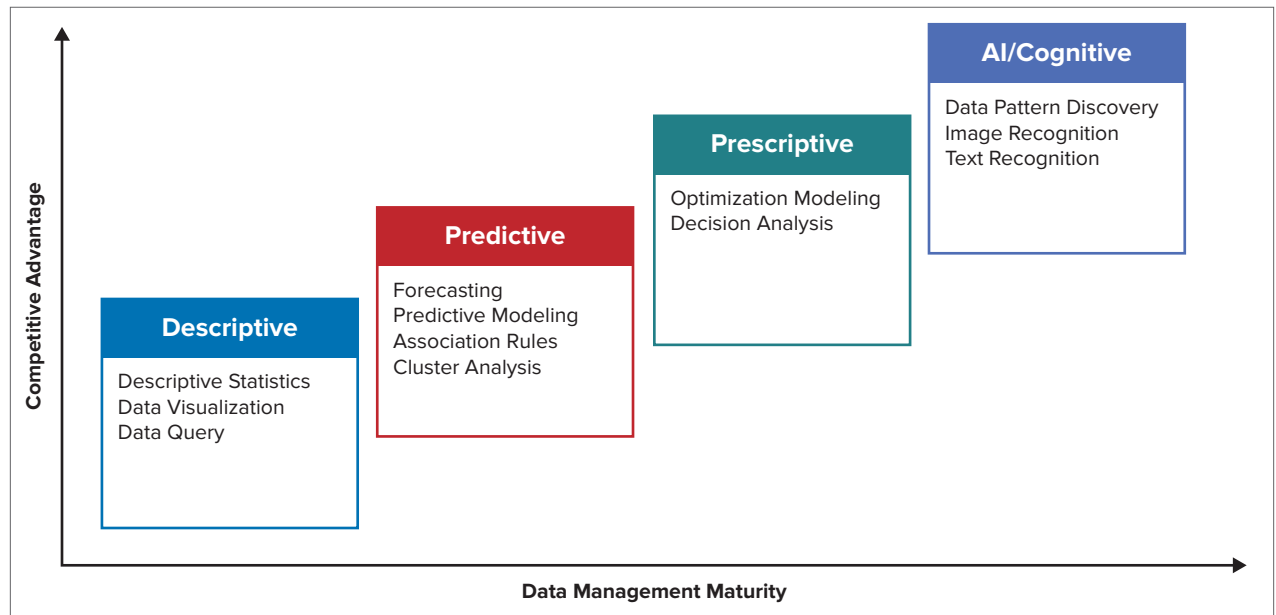
in a situation described as “data rich but information poor.” Until recently, many organizations were making decisions based upon intuition or opinion versus data-driven knowledge. Data analytics techniques provide an excellent opportunity to bridge the gap between information and insights.

As technology continues to improve and dominate innovative processes, analytics will become a ubiquitous part of everything we do. To prepare you for this, we explain how to creatively approach a problem, comprehend the essence of communication and collaboration, understand key elements of project management, and complete a successful project. These skills are the most critical in the age of data analytics.

Analytics Levels and Their Impact on Competitive Advantage

Analytics involves techniques as simple as descriptive statistics and visualization, as well as more advanced predictive modeling, prescriptive, and newly emerging artificial intelligence (AI) and cognitive analytics. As organizations adopt more advanced techniques (i.e., predictive, prescriptive, and AI methods), higher data management maturity is required to achieve a competitive advantage, as depicted in Exhibit 1-2.

Exhibit 1-2 The Competitive Advantage of Marketing Analytics



Source: Adapted from SAS.

Descriptive analytics are a set of techniques used to explain or quantify the past. Several examples of descriptive analytics include: data queries, visual reports, and descriptive statistics (e.g., mean, mode, median, variance, standard deviation). This type of information is essential to summarize questions related to how many and how often situations occur. For example, how many customers use a mobile app each day, and how often do they visit a website within the same month? Customer needs and motivations are not always understood, but these fundamental insights provide a foundation for marketers to explore what is fueling the behavior. Descriptive analysis can be especially helpful when marketers collect data from a survey. Retailers such as Dick’s Sporting Goods and Dunkin’ survey customers in return for a coupon code or free product following a recent experience. The survey questions focus on capturing whether customers feel stores maintained correct levels of stock or were satisfied with the purchase experience. Overall averages and trends

resulting from this technique can be beneficial in reinforcing existing practices and determining how the company might improve the customer's experience moving forward.

Predictive analytics is used to build models based on the past to explain the future. Mathematical models examine historical data to predict new values, needs, and opportunities. For example, historical customer sales data can be used to predict future sales. You might recall when Target predicted a teen girl was pregnant before her father was aware.⁵ How does this happen? Target Corporation collected data from customer purchases and used predictive modeling to classify customers as “pregnant” or “not pregnant.” Customers identified as pregnant based on their purchases were then sent sales promotions for early pregnancy to capitalize on a significant revenue stream. Zillow, an online real estate database and analytics platform, also develops predictive models from publicly available city housing data. Zillow predicts a “Zestimate” or value for every home based on over 100 predictors for each of the 100 million homes in its database.⁶

Prescriptive analytics identifies the best optimal course of action or decision. Consider, for example, how UPS efficiently maps drivers through a city using optimized routes that reduce the number of left turns, or how airlines maintain productivity, reduce costs, and increase customer satisfaction by optimizing flight and crew scheduling. Price optimization, also a growing e-commerce practice, is used by Amazon to develop pricing strategies and remain competitive in the retail industry. In fact, the company has reported changing prices more than 2.5 million times a day⁷ to influence customer behavior and maximize revenue. Kellogg Company is the world's leading cereal company and second-largest producer of cookies in the world, with annual revenues exceeding \$14 billion. In the cereal business alone, Kellogg Company has more than 90 production lines and 180 packaging lines. This requires tremendous coordination to meet customers' demand at a low cost. Kellogg uses optimization models to forecast sales and determine what should be produced and shipped on a daily basis. The company also uses optimization modeling to improve supply chain infrastructure. To do this, Kellogg must identify the number and location of plants to produce the required level of production that minimizes excess capacity and provides the inventory to meet customer demand. Using this modeling, Kellogg estimates a savings of over \$475 million a year.⁸

Artificial intelligence (AI) and **cognitive analytics** are designed to mimic human-like intelligence for certain tasks, such as discovering patterns in data, recognizing objects from an image, understanding the meaning of text, and processing voice commands. Artificial intelligence (AI) and cognitive analytics use machine learning to understand new data and patterns that have never been identified. **Machine learning** is the development of algorithms and statistical models that allow computers to learn and improve from experience, without being explicitly programmed. This method produces tasks that are often beyond the reach of a human in speed and accuracy. The techniques “learn” over time by updating the algorithm as new information becomes available.

Using technology powered by AI, Olay, a Procter & Gamble skincare line, has doubled its sales conversion rate.⁹ Olay encourages customers to upload personal photos to an app where it uses image recognition to identify the age of skin and then recommends specific products. This process can be completed in the convenience and privacy of a customer's home on their computer or mobile device, so customers know what they want before arriving at the store. Mtaylor, an online clothing retailer, also uses customer images to create a customized clothing fit. The customer uses an app that measures 17 different points to develop a personalized fit and recommended size. In fact, Mtaylor claims their method is more accurate than a professional tailor. In the case of Olay and Mtaylor, AI engages the customer to obtain data and then produce personalized recommendations for products.

Indeed, today's AI technology enables almost any company to augment or complement human capabilities in developing customer solutions. Hitachi, a Japanese multinational

PRACTITIONER CORNER

Theresa Kushner | Advisory Board Member at data.world and Partner at Business Data Leadership



Ian Tuttle

Theresa Kushner, partner in Business Data Leadership, comes with over 20 years of experience in deploying predictive analytics at IBM, Cisco, VMware, and Dell. She is an accomplished, Business-Centric Executive and Board Advisor who understands data and leads companies through transformations in the midst of rapid technology, regulatory issues, and market disruptions. Theresa has expertise harnessing data analytics and company and customer information to lower costs and contribute multi-billion dollar growth for publicly traded, technology leaders. She has co-authored two books: *Managing Your Business Data from Chaos to Confidence* and *B2B Data Driven Marketing: Sources, Uses, Results*.

Q *The Wall Street Journal* recently reported that automotive companies were struggling with the use of artificial intelligence.¹⁰ Barriers to adoption included “difficulties in implementation technology, limited data for algorithmic training, a shortage of talent in data science, and uncertainty about the return on investment.” Theresa, from your perspective, how can companies address these problems prior to pursuing AI?

A To begin with, I think that all traditional companies are having trouble making the transition to an environment where AI and machine learning are applied effortlessly because they haven’t relooked at their strategy and made room for this kind of transformation. For example, an automotive company obtains over 8 billion records a day from its automobiles, but what to do with those records is not always thoughtfully planned against a strategy. Let’s just assume that someone in engineering thinks it’s a smart idea to track whenever a seat belt is buckled. That immediately helps with safety

standards and you can get some great data off that click. However, the remainder of the engineering staff may be looking at several other events that could be just as telling about the safety of the car and driver. Are those events planned with the seatbelt click or without considering it at all?

The holistic approach of data management and how it relates to a strategy is very difficult for companies like GM or any of the car manufacturers, because they did not begin with the end in mind. They are improvising the use of AI as they go along. All the problems mentioned—such as difficulties in implementation, limited data for algorithmic training, and even the shortage of trained personnel—can be traced back to a company not articulating how AI/Machine Learning supports their strategy. If the strategy were there, the companies had a strategy connected to their data, and they would have no problem articulating a return on investment.

Continued to next page

conglomerate, is using AI named “H” to discover patterns that typically go undetected by humans. The H process generates customer solutions and selects the best options to improve operations at call centers, retail sales, financing alternatives, warehouse management, and similar tasks.¹¹ Applications like “H” can easily automate and improve customer interactions to increase sales and, ultimately, customer loyalty.

1.2 Defining the Right Business Problems

Marketing analysts face complex challenges in today’s data-intensive and competitive business environment. They are often confronted with multiple courses of action that must be completed quickly. Evaluating these alternatives and choosing the best action

forward is at the heart of decision analysis. But one of the most important initial steps in the marketing analytics journey is defining the right business problem to pursue. A successful understanding of business problems requires deep knowledge of the customers' journey, from how they search to where they purchase, and how satisfied they are with products and services. Problem identification helps to uncover strategic business opportunities. Business initiatives that improve market share, establish a better relationship with the customer, or position the enterprise to take advantage of innovation are a few business strategies that can be supported by an analytical approach.

One of the most critical steps in the analytics process is to start with an understanding of the real business question the data needs to address. Knowing the right business question leads to smarter decisions that drive impact.

Continued from previous page

PRACTITIONER CORNER

Theresa Kushner | Advisory Board Member at data.world and Partner at Business Data Leadership

Q What are the greatest challenges facing analysts when attempting to define the business problem?

A The biggest challenge facing analysts when attempting to define the business problem is their innate way of thinking. Analysts have a very structured way of thinking about problems, but business problems are not always structured, especially in marketing and sales, where relationships play a part in the success of most actions. The challenge usually begins when the teams are defining the problem, and it begins at a very basic level. Ensuring that everyone understands what needs to happen, how it will happen, who will be responsible for it happening—these are key decisions that must be made by all involved in the analysis *before* it begins.

Q What is the impact of inaccurately defining the business problem prior to undertaking an analytics project?


A I've seen too many "problems" given to analysts to analyze that weren't carefully thought through. For example, a marketing manager commissions an algorithm to predict which customers will buy the new product to be made available next quarter. He doesn't tell the analyst how the information will be used. He just wants a "list." The analyst assumes that the "list" will be used with the inside sales team and develops a ranked list of customers that might be in the market. But the marketing manager really wanted a "list" of potential customers who might be willing to beta test the new product. The difference between success and failure of this analytic project depends on how thoroughly the problem is discussed and how vetted the solution is for its applicability to the problem.

Continued to page 13

How do you arrive at the right business problem? The process begins by understanding the intent and business considerations behind the question. Consider you are working at a retailer that was once the largest in the country, but it recently filed for bankruptcy. The marketing executive calls you to her office and explains that large investments were made specifically to develop mobile applications. Unfortunately, visitors are registering for an account, but then not using it for purchases. The executive asks you to determine how to entice those first-time users back to the mobile application, because the company was relying upon this technology to make them competitive again. At this point, is the underlying business problem evident? A better understanding of the business problem can be gained through interviews with business stakeholders. Conversations that more

broadly discuss problems facing the company are likely to uncover more relevant issues. During your initial conversation with the marketing executive, she explains that the threat of competition from Amazon, Target, and Walmart continues to rise and profits are falling. Moreover, the company’s brick-and-mortar locations are outdated, rely on obsolete technology, and are experiencing declining foot traffic and low levels of customer satisfaction. It is apparent the company should not limit the investigation to attracting new customers, and should expand it to encompass how to retain loyal customers. If you were to proceed with the original project and focus only on returning visitors to the mobile application, it would mean overlooking important relationships, and stakeholders would criticize your limited analytics approach. Thus, marketers must incorporate relevant stakeholder inputs through discovery methods to collectively understand the business problem and align projects to achieve business objectives.

MARKETING ANALYTICS EXTRAS



A good analytics approach must engage stakeholders in determining project requirements. Marketing analysts must seek cooperation from appropriate stakeholders in the project outcome. These stakeholders may include customers, employees, suppliers, and subject matter experts. Collection of feedback related to the business problem could occur via interviews, observation, surveys, and brainstorming sessions.

Discovery begins in asking the traditional six discovery questions: What, who, where, when, why, and how. Exhibit 1-3 provides samples of discovery questions and how they might be useful when considering the business problem.

Exhibit 1-3 Asking the Right Questions to Identify the Right Business Problem

BUSINESS CONSIDERATION	SAMPLE QUESTION
Context	What happened? What is the current problem we are trying to solve? What is the potential opportunity? Why is there an interest in solving this particular problem? What is the business doing to mitigate or solve the problem? What efforts have been made in the past? How has this problem evolved over time?
Impacted unit	Where did this problem happen? What divisions are impacted by this problem? When did it take place?
Root-cause analysis	What might have caused this? What do you think continues to drive this problem?
Timeline	When do decisions need to be made? What is the optimal timeline for reaching milestones along the way?

Continued from previous page

Stakeholder	<p>Who is asking for the analysis?</p> <p>Who are the executives interested in the results of the analysis?</p> <p>Who will be impacted by the analysis and subsequent recommendations?</p> <p>Who will carry out the analysis?</p> <p>What financial or emotional interest is involved from stakeholders? Is it positive or negative?</p>
Expected impact	<p>What are the actions to take based on the analysis?</p> <p>What support will end users have?</p> <p>What is the anticipated ROI from solving this problem?</p> <p>What are the ethical implications of the analysis?</p>

Source: Adapted from Tony de Bree, "8 Questions Every Business Analyst Should Ask," *Modern Analyst*, <http://www.modernanalyst.com/Resources/Articles/tabid/115/ID/179/8-Questions-EveryBusiness-Analyst-Should-Ask.aspx>; and Piyanka Jain and Puneet Sharma. *Behind every good decision: How anyone can use business analytics to turn data into profitable insight* (AMACOM, 2014).

In an effort to define the right business problem, it can be useful to follow the SMART analytics principles. The **SMART principles** can be used as a goal-setting technique.¹² The acronym stands for specific, measurable, attainable, relevant, and timely (see Exhibit 1-4). First, the project's goals should be specific and clearly defined. Second, the project should be trackable and the outcomes measurable. For example, One SMART analytics goal could be to determine changes to the mobile application that will most efficiently increase returning visitors by 10 percent on a quarter-by-quarter basis compared to the same quarter last year. If data related to returning mobile application visitors is unavailable, it would be necessary to develop a project to obtain the data. The new goal would reflect the data acquisition and be stated as follows: "By the end of 6 months after the mobile application data has been collected, the data will be analyzed to determine the most efficient app changes to increase returning visitors by 10 percent on a quarter-by-quarter basis compared to the same quarter last year." Third, project goals should be reasonable to achieve. Fourth, the project should solve the analytics problem and align with the business objectives. Fifth, the project should be completed in a timely manner. Developing sound goals and objectives allows the analyst to monitor the project's progress, ensure it remains on track, gain visibility among stakeholders, and verify that everyone is on the same page.

Exhibit 1-4 SMART Principles

<p>S</p> <p>Specific</p> <p>The goal should be clearly defined.</p>	<p>M</p> <p>Measurable</p> <p>Progress of the goal should be trackable and have a measurable outcome.</p>	<p>A</p> <p>Achievable</p> <p>The goal should be reasonable to accomplish.</p>	<p>R</p> <p>Relevant</p> <p>The goals should solve the analytics problem and align with business objectives.</p>	<p>T</p> <p>Timely</p> <p>A timeframe to successfully complete the analytics project should be determined.</p>
---	---	--	--	--

Following the SMART analytics goal-setting technique is important. But equally important is examining the potential success of the analytics project and whether it makes a valuable impact. To do so, the opinions of the most powerful stakeholders should be included when developing project goals and success measures, as well as in evaluating the results.

When the SMART analytics goals are identified, it is time to focus on understanding the data requirements. Let's begin by taking a closer look at data sources.

1.3 Data Sources

Data sources consist of both primary and secondary data. **Primary data** is collected for a specific purpose. Companies conduct surveys, focus groups, interviews, observations, and experiments to address problems or answer distinct questions. For instance, Walmart and other companies are now observing customers in real-time through facial recognition software. The objective is to detect customers that are unhappy or need further assistance while shopping and, in turn, have an employee respond to their needs.¹³ These observations provide primary data to achieve a specific objective.

In contrast, **secondary data** relies on existing data that has been collected for another purpose. While secondary data might not address specific or current problems, it could be useful in formulating ideas about how to ask the right questions or to design future data collection initiatives. At the same time, internal and external secondary data sources can be useful in exploring current business questions. Sources of secondary data include:

- *Public datasets:* Google launched Google Dataset Search in 2018 to enable scientists, analysts, and data enthusiasts to find data that is important for their work. Each dataset has a description and a discussion of the problem the dataset can address. Google Dataset Search includes data from NASA, NOAA, Harvard's Dataverse, GitHub, Kaggle, and other sources.
- *Online sites:* Online browsing behavior, purchase history, and social media chatter have become increasingly popular sources of data. As one example, food safety is a common concern in the restaurant industry. The third-largest fast-food chain in the United States, Chick-fil-A, is now exploring social media content to identify words or phrases often associated with food safety issues.¹⁴ This data and analytics are used to produce results that are quickly made available to managers on a corporate dashboard to make local decisions. Managers also review the results of website postings and identify individual customers to contact. The objective is to eventually make selected information available on GitHub, a software development platform where creators can share programming insights at no cost.
- *Mobile data:* Most mobile applications track data so companies create more effective marketing strategies. As consumers have increasingly adopted mobile applications, companies like restaurants and clothing retailers have developed mobile apps that record customer purchase behaviors and geographic locations. Aki Technologies, for example, a mobile advertising company, offers a platform that companies can use to track mobile phone activity and then develop customer segments based on certain behaviors.¹⁵ The data is then used to engage customers through personalized advertising campaigns or messages that increase store traffic.
- *Channel partners:* Multiple companies often operate within a distribution channel. The companies, referred to as channel partners, include suppliers, wholesalers, distributors, or retailers. Each member of the channel collects data unique to their business, but the data frequently provides value to other partners in the channel. In the fast-paced retail environment, where customers can purchase and receive

products through same-day delivery services, brick-and-mortar retailers are increasingly concerned with maintaining the correct levels of inventory. Walmart strives to meet the needs and wants of customers by collecting data that enables them to always have products in stock for purchase. To reduce the time involved in restocking products, they are now sharing inventory data with suppliers and other channel partners.¹⁶ They expect suppliers will be better prepared to restock products at the right time at the right location.

- *Commercial brokers:* Companies collecting and selling both public and private data to a wide range of customers have emerged in recent years. Acxiom is one example of a consumer data broker. In a typical year, the company aggregates and sells over 10,000 types of data, including socioeconomic status, health interests, and political views on more than 2.5 billion consumers.¹⁷ Companies purchase this data to create customer profiles that can be used to target a wide variety of target segments.
- *Corporate information:* In this era of big data, many companies constantly collect and store data ranging from general business transactions to customer exchanges with accounting, finance, sales, customer service, and marketing. For example, customer service interactions are useful when the marketing department wants to better understand product quality, customer satisfaction, and loyalty. Integrating data across functional areas enables companies to better understand customer interactions based on a more holistic view of transactions.
- *Government sources:* This is an important source of secondary data collected by local, state, and federal government agencies. More than 200,000 datasets are searchable by topic on Data.gov (see Exhibit 1-5), including the following data sources, which are directly applicable to marketing analysts:
 - The U.S. Census Bureau is part of the U.S. Department of Commerce data. It includes data related to the population, economy, housing, and geography.
 - Consumer complaint data provides customer sentiments about financial products and services.
 - Demographic statistics by ZIP code, gender, ethnicity, and citizenship.
 - Fruit and vegetable prices for over 153 commonly consumed products are available from the Department of Agriculture.
 - ZIP code data showing tax return data by state and ZIP code level.

Exhibit 1-5 U.S. Government's Dataset Topics



Source: Data.gov.

1.4 Data Types

Types of Data

Data is facts and figures collected, organized, and presented for analysis and interpretation. Data is available in two main forms: structured and unstructured.

Structured Data Structured data is made up of records that are organized in rows and columns and are easily searchable and analyzable using computer algorithms. This type of data can be stored in a database or spreadsheet format. It includes numbers, dates, and text strings that are stored in a clearly defined structure. The data is easy to access and analyze using descriptive, predictive, prescriptive, and AI data analytics techniques.

Unstructured Data Unstructured data includes text, images, videos, and sensor data. The data does not have a predefined structure and does not fit well into a table format (within rows and columns). Examples of this type of data include voice recording from customer service calls, text, images, video recording, social media conversations, and the Internet of Things sensor data. Unstructured data could benefit from advanced analytics techniques such as AI to prepare and analyze. When possible, unstructured data is converted to a structured format prior to analysis. The number of companies collecting unstructured data has increased substantially as technology has advanced to efficiently support manipulation and exploration of this data type.

Both structured and unstructured data are important in executing marketing analytics. As noted, structured and unstructured data come in different formats and measurement types.

Continued from page 8

PRACTITIONER CORNER

Theresa Kushner | Advisory Board Member at data.world and Partner at Business Data Leadership

Q How are companies using both unstructured and structured data in solving business problems?

A Companies started combining unstructured and structured data to better understand their customers. Companies like Cisco Systems and Dell Technologies combine unstructured data from conversations on their support sites with structured data about their individual customer accounts.

Q What technologies have facilitated the integration and use of both data types?

A Graph databases have greatly improved the combination of these kinds of data structures. But users who understand how to implement these technologies are still lagging.

Q How do you see this evolving over the next 5 years?

A In the next 5 years, we should see more and more of the combination of these data structures aided by AI and machine learning.

Continued to page 21

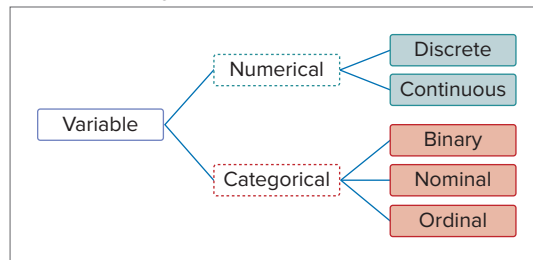
Data Measurement

Data measurement can be categorized in a variety of ways based on the type and means for collection (see Exhibit 1-6). The two main types of data measurement most often explored in the remaining chapters are numerical and categorical (see Exhibit 1-7).

Exhibit 1-6 Data Measurement Definitions and Examples

DATA MEASUREMENT TYPE	DEFINITION	EXAMPLE
Discrete	Is measured as whole numbers: 1, 2, 3, . . .	Number of items purchased on a website
Continuous	Includes values with decimals: 1, 1.4, 2, 2.5, 3.75, . . .	Amount of time spent on a website
Binary	Has only two values	Yes/No, True/False
Nominal	Consists of characteristics that have no meaningful order	Marital Status, Country of Origin
Ordinal	Represents rank order	Ranking products/services in order of preference
Interval	Has fixed interval between data points	Degrees Fahrenheit
Ratio	Has a true zero point	Product Sales, Age

Exhibit 1-7 Types of Data Measurement



Numerical Data are considered quantitative if numeric and arithmetic operations can be applied. For instance, sales data or visits to websites are numerical because they can be summed or averaged. Numerical data can be either *discrete* (**integer**) or *continuous* in nature. **Discrete data** is measured as whole numbers: 1, 2, 3, . . . The number of items purchased by a customer on a retailer’s website is discrete. In contrast, **continuous data** can include values with decimals: 1, 1.4, 2, 2.5, 3.75, . . . The amount of time a customer spends on a retailer’s website would be continuous.

Categorical **Categorical data** exist when values are selected from a group of categories. A common example might be marital status. You may notice, categorical data can only be summarized by calculating the proportion and count of occurrences across and within categories.

Categorical variables can be one of three types: binary, nominal, or ordinal. **Binary** categorical data can have two values—for example, yes or no. This can be represented in different ways such as 1 or 0 or “True” and “False.” Binary data is commonly used for classification in predictive modeling. Examples of binary variables include whether a person has purchased a product or not, or uses Twitter or not.

Nominal categorical data consist of characteristics that have no meaningful order. Marketers might inquire about the customer’s country or marital status. There is no

magnitude of value because a person cannot be half married and each category's characteristics are equally meaningful. The characteristics reflect the state of being: United States, China, Russia, Saudi Arabia, Mexico, United Kingdom, France, Germany, married, unmarried, divorced, widowed, and so on.

On the other hand, **ordinal** categorical data represent meaningful values. They have a natural order, but the intervals between scale points may be uneven (e.g., the rank order from the top product to the second may be large, but the interval from the second-ranked product to the third-ranked may be small). Customers might respond to a question such as do you prefer this brand more than or less than another brand. Another example might be when a company asks customers to rank products in order of preference.

Categorical variables require special consideration in preparation for modeling. How to prepare these variables will be discussed in a later chapter.

Metric Measurement Scales

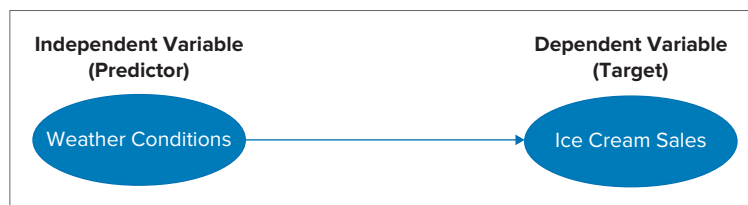
Scales can also be metric. Metric scales can be measured as **intervals** or **ratios**. Both of these scales possess meaningful, constant units of measure, and the distance between each point on the scale are equal. However, there is a difference between these scales. Interval variables do not include an absolute zero. When thinking about weather, 0 degrees Fahrenheit means nothing, except that it is very cold. But ratio scales have an absolute zero point and can be discussed in terms of multiples when comparing one point to another. Product sales of \$0 means nothing and sales of \$100 is twice as much as \$50. Similarly, zero indicates a lack of any weight, and 50 pounds is half of 100 pounds.

1.5 Predictors versus Target Variable

Types of Variables

Variables are characteristics or features that pertain to a person, place, or object. Marketing analysts explore relationships between variables to improve decision making. Consider a simple example. An analyst is investigating the relationship between two variables: weather conditions and ice cream sales. Does the weather impact customer ice cream purchases? Weather conditions would be considered the **independent variable** or what influences or drives the **dependent, target, or outcome variable** (ice cream sales). Warmer weather increases the likelihood that more ice cream will be purchased (see Exhibit 1-8). What other variables might impact ice cream sales? Although the example in Exhibit 1-8 only uses two variables, companies often use multiple variables at the same time as inputs to systems that process data and use it to predict dependent variables.

Exhibit 1-8 Example of Variable Types



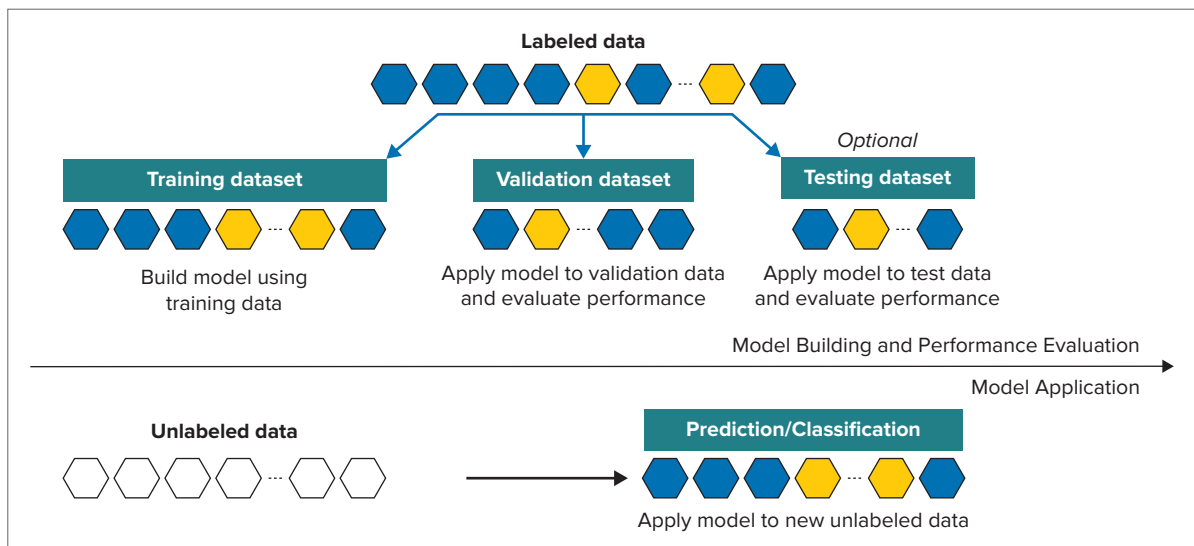
In practical marketing analytics applications, variables are designated as independent (predictor) variables or dependent (target) variables.

1.6 Modeling Types: Supervised Learning versus Unsupervised Learning

Depending on the nature of the business problem being addressed, different types of algorithms can be used. In this book, we will focus on two types: *supervised learning* and *unsupervised learning*. **Supervised learning** suggests that the target variable of interest is known (e.g., ice cream sales; click or no click) and is available in a historical dataset.

The historical dataset can also be referred to as labeled data (because the target variable is available) and is divided into a training dataset, a validation dataset, and an optional testing dataset (also known as a holdout sample). The **training dataset** is the data used to build the algorithm and “learn” the relationship between the predictors and the target variable. The resulting algorithm is then applied to the **validation dataset** to assess how well it estimates the target variable, and to select the model that most accurately predicts the target value of interest. If many different algorithms are being compared, then it is recommended that a third data called **testing dataset** be used to evaluate the final selected algorithm and see how well it performs on a third dataset. The final selected algorithm is then applied to predict the target variable using new unlabeled data where the outcomes are not known, as shown in Exhibit 1-9.

Exhibit 1-9 Supervised Learning Steps



When the target variable is continuous, supervised learning is referred to as *prediction*. Let's say a retail company wants to understand customer buying behavior, specifically the purchase amount to create personalized offers. We will need to build a model to predict the purchase amount of customers against various products using labeled data (i.e., historical data that includes how much the customer spent on each product). The model can then be used to predict the purchase amount of new customers with similar characteristics.

When the target variable is categorical (typically binary—buy/no buy), supervised learning is called *classification*. Consider a large U.S. bank whose objective is to acquire new credit cardholders based on a special promotion. The historical data includes records of customers who have qualified (and not qualified) for past credit card offers after receiving the special promotion. An algorithm would be trained and validated on labeled data and then used to predict who should be targeted in the new promotional campaign. Exhibit 1-10 shows other examples of supervised learning applications using a variety of predictors.

Exhibit 1-10 Examples of Supervised Learning Applications

PREDICTOR (X)	TARGET (Y)	APPLICATION
Purchase histories	Future purchase behavior	Customer retention
Store transaction details	Is the transaction fraudulent?	Fraud detection
Faces	Names	Face recognition

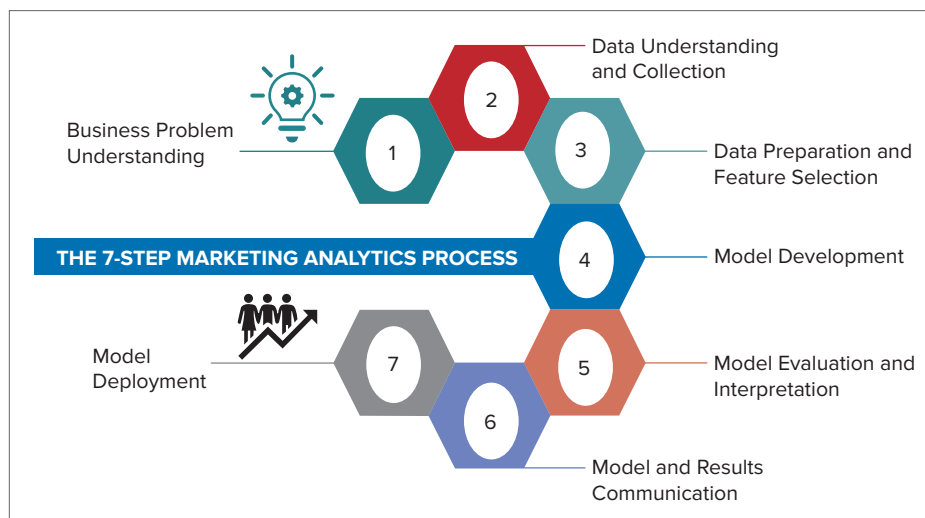
Unsupervised learning has no previously defined target variable. The goal of unsupervised learning is to model the underlying structure and distribution in the data to discover and confirm patterns in the data. Examples of these techniques include association analysis and collaborative filtering. Sephora runs for customers on its website using the “you may also like product X” based on an assortment of past purchases, or when Amazon indicates “others who bought this item also bought products X, Y and Z.” Sephora and other companies also use cluster analysis to group customers into homogenous sets based on their loyalty (high, medium, and low) using their purchase history, the amount spent each year, and other key demographic and purchasing related variables.¹⁸

Supervised and unsupervised learning can be used together to gain more insights. For example, after conducting the unsupervised learning to determine customer loyalty segments, supervised learning can be used to predict purchase amounts for each of these segments. Whether the algorithm is supervised or unsupervised, the modeling process must be developed to represent how real-world problems begin, starting with a business problem and working toward a solution that makes a business impact. Modeling steps are discussed in greater detail in the following section.

1.7 The 7-Step Marketing Analytics Process

There are seven steps involved in the marketing analytics process (see Exhibit 1-11). Data modeling is only part of the journey, not the full marketing analytics journey. The 7-step marketing analytics process is iterative and continuously evolves to develop and manage improvements in the marketing analytics cycle. Each step plays an important role in achieving a successful outcome.

Exhibit 1-11 The 7-Step Marketing Analytics Process



Step 1: Business Problem Understanding

Most marketing analytics models are developed when a business identifies a problem. The company may be experiencing low website traffic or conversion, high customer churn, or new product or market growth is slower than expected. The idea is to develop a model using analytics to understand the problem better and design a solution. One of the key elements in this step is to question whether the problem the business is presenting is, in fact, the correct problem, as was discussed in section 1.2. Due to today's dynamic business environment, companies frequently overlook or do not recognize problems. To avoid this situation, several questions should be asked:

- Exactly what are you trying to understand and solve?
- How will the stakeholder(s) use the results?
- Who will be affected by the results?
- Is this a single, short-term problem or an ongoing situation?

These initial questions are essential to make sure the right problem is being addressed.

MARKETING ANALYTICS EXTRAS



Most experienced analysts recommend spending time thinking about defining the marketing analytics problem and determining its scope and feasibility before starting the next step. Given that most real-world problems are complex and often broad in scope, this recommendation is essential.

Step 2: Data Understanding and Collection

There are many different sources of data within an organization. The marketing analysts first job is to identify where the data is stored, its format, and how it can be combined to understand the question at hand. This step typically includes examining databases inside and outside the organization, and then talking with different key data owners and stakeholders (for example, the customer relationship manager, IT manager, or sales manager). It also includes observing and understanding organizational processes to determine if the problem identified is the actual problem or if it is a symptom of another underlying problem.

Once a better understanding of the problem is established, the analyst typically samples data from the selected databases to obtain records for the analysis. For example, the marketing analyst may use SQL code (a type of programming language introduced in Chapter 2) to examine past purchases and returns of customers. In various Practitioner Corners throughout the book, you will learn how marketing analysts collect, clean, and prepare data for analysis. These are basic tasks in the data understanding process of the marketing analytics cycle.

Marketing analysts must have a good understanding of the types and sources of data.

It is important to understand the data prior to analysis.

What if Apple failed to test the strength of iPhone screens when they are dropped, or Tyson did not keep track of its chicken products so salmonella outbreaks could be

tracked and resolved? Starting with all the information necessary in identifying, understanding, and solving a problem is clearly critical to finding the correct solution and optimizing the marketing analytics process.

Step 3: Data Preparation and Feature Selection

Data in different formats is combined in this step. To do so, the unit of analysis (e.g., customer, transaction, subscription plan) and the target and predictor variables are identified. The data columns (features) of the target and predictor variables are then visually and statistically examined. For example, scatterplots showing the relationship between the continuous target and each of the continuous predictors can be useful to identify patterns in this relationship. Data is cleaned by identifying and determining how to deal with missing values, data errors, or outliers. The data from different data sources describing the unit of analysis is also merged so data from both sources is measured consistently and can be used in developing the models.

Other features are further refined in this step. For example, dates might be adjusted to represent the day of the week, week, month, or year. In addition, predictors that have a strong relationship with the target variable and are not highly correlated with each other (predictors that are unique and not highly related) are included in the analysis to improve the reliability and accuracy of the predictive model. In this step, predictors might be eliminated, but they could also be transformed to improve the measurement. For example, the focus of the problem may be mobile phones priced less than or equal to \$200 and those greater than \$200. Rather than examining continuous monetary values, the feature can be changed to a binary variable ($= < \$200$ or $> \$200$). Similarly, if the research involves company size and performance, the focus could be companies with fewer than or equal to 500 employees and more than 500 employees. Understanding the meaning of each variable and its unit of analysis is an essential task in this step.

Step 4: Modeling Development

Steps 1 through 3 represent about 80 percent of the analyst's time, but serve as an important foundation for the rest of the steps in the process.¹⁹ In step 4, the analyst uses analytical skills. A good model is one that represents the real-problem accurately and includes all key assumptions.

In this step, the analyst selects the method to use. The choice depends on the target variable type and availability and the business question addressed. The possible options include classification and prediction when a target variable is defined, and clustering or association when no target variable is available. The possible options are classification, prediction, clustering, or association. If the problem is supervised, the analyst will need to partition the data into three parts as previously indicated: training, validation, and test datasets. The analyst will also have to decide on appropriate modeling techniques such as regression and neural network, which are explained in the following chapters. More than one modeling technique is typically used in this process, and each includes a variety of features. Different models should be tried to identify the one that provides the best accuracy, speed, and quality.

A key idea to remember is that the model should be simple, practical, and useful. Netflix paid a million dollars for a model it never used due to its complexity.²⁰ Until recently, some of the most commonly encountered analytics problems were solved using simple techniques such as decision trees and regression analysis. The results were not necessarily as accurate, but the techniques were simple to understand and apply, and useful solutions could be developed.

Step 5: Model Evaluation and Interpretation

This step ensures the modeling is accurately performed and provides the best predictions for future applications. The model is evaluated to identify the algorithm providing the best solution. Initially, the algorithm is run on the validation dataset to determine how well it will predict the relevant target variable (dependent variable). If the validation results show high accuracy, then the model can be recommended to predict new cases and address the business problem. In some instances, the top model can be evaluated using an optional testing dataset to assess how well the final selected model might perform on new data.

Step 6: Model and Results Communication

The modeling step provides a set of recommendations. Understanding differences in the perspectives of problem-solving skills is critical because most people may not have a clear understanding of the modeling techniques used. It is key, therefore, for the analyst to present the model in a way that other people can understand, particularly management. Otherwise, the model may never be approved for implementation.

A good approach for this step is to collaborate with key stakeholders early in the process. If key stakeholders such as executives and managers have been involved in providing feedback from the beginning of the process (e.g., providing data and evaluating the progress), they are more likely to understand and support the recommended modeling approach.

A full understanding of the model is another important consideration. Whether the model is simple or complex, it should be explainable in straightforward terms with the appropriate visualization of results. For example, managers appreciate a regression model that includes a clear representation of the relationship between the target and predictors and that rapidly guides them in determining if their initial questions were answered.

Step 7: Model Deployment

The model completion and execution step is not finished until it has been implemented and is running on real-time records to offer decisions or actions. For example, a web recommender system is not considered complete until the system is used to make recommendations to customers during online purchases. Model implementation is typically approved by management for deployment, but only after full buy-in will the model add real value in making better decisions. Typically, this step involves other key stakeholders such as IT specialists, customer service representatives, or the sales team. These individuals should also train on implementing the system to ensure they understand how the model is executed and applied.

A key consideration throughout the 7-step marketing analytics modeling process is to evaluate the ethical dimensions of the analysis. Are the privacy and anonymity of the subjects being protected? Does a bias exist in the data that could impact the analytics results? Are the model results accurate? If not, some subjects may be misclassified and have a negative effect on the results. For example, applicants could be denied a bank loan, or purchase conversions may not increase as expected. At times, the model may be correct, but the objective is unfair to some subjects or unrealistic in its predictions. IBM, Microsoft, Facebook, Google, and other companies have created analytics review boards and ethical codes to evaluate the fairness of the analytics modeling. Another

issue to keep in mind is that the data, features, data cleaning, and the model are determined by analysts. Thus, ethical training, ethical codes, and clear guidelines should be established and communicated to everyone working on developing the analytical model.

Continued from page 13

PRACTITIONER CORNER

Theresa Kushner | Advisory Board Member at data.world and Partner at Business Data Leadership

Q What are the ethical considerations that marketing analytics students should pay attention to as they begin their career in analytics?

A The one ethical consideration that marketing analytics students need to pay close attention to is bias. This takes various forms. There is bias in the data collected for analytics projects. There is bias in training data for AI/Machine Learning projects. There is bias in applying the learning from the analytics to the business problem. Not all bias is bad, but knowing that you have it and that it must be managed is something that most students in marketing analytics do not recognize. The way to avoid having bias enter the algorithms or the data

is to make sure you have checkpoints throughout the process that look for bias. Most data science processes do not consider this aspect. The other way to minimize bias is to create diverse teams that do the analytics. This doesn't necessarily mean diversity in race or gender, but it does mean diversity in thought. People who approach problems from different perspectives are a requirement in today's data science teams. Left-brained and right-brained people are needed. In fact, the most recent needed addition to a data scientist team is a data archeologist, someone who understands how data has been curated in the past—someone who uses both right- and left-brain thinking.

Continued to next page

1.8 Ethical Considerations

As AI and machine learning use become increasingly integrated into marketing applications, it's becoming more important than ever to pay attention to ethical considerations. While advanced analytics has the potential to provide valuable insights into consumer behavior, there is a significant risk that they could be used in ways that harm customers. For example, sensitive information about customers could be collected without their consent, customer can be manipulated to make a purchase they wouldn't make otherwise, or profiles could be created and evaluated based on factors like race, gender, or religion. In 2019, Facebook was sued for enabling advertisers to exclude certain racial, ethnic, and religious groups from seeing their ads during problem framing and data collection.

These activities could lead to serious harm to the customers including financial losses, reputation damages, or even bias and discrimination. Furthermore, unethical machine learning algorithm could damage customer's trust in the brand and businesses that use it. The literature shows that consumers are concerned more than ever about how companies are collecting and using their personal data, they will be likely to disengage with brands that use these technologies in harmful ways.

Companies need to pay attention to ethical considerations of the use of AI and machine learning in marketing applications. This involves being transparent about

how data is being collected and used, obtaining consent from consumers before collecting their data, and ensuring that algorithms are not being used to bias or discriminate against individuals. By ensuring the ethical use of machine learning, companies can ensure that their marketing applications are used to protect their customers from harm, preserve their trust in technology, and their relationship with the brand.

Given the importance of this topic, we will address ethical considerations throughout the book. Understanding bias in AI can help you become a more informed and effective marketer and help you advance your career. This knowledge is also essential to help you create marketing campaigns and applications that are equitable and just for our society.

1.9 Setting Yourself Apart

Marketing analytics experts work on solving complex and important problems. Regardless of which area in marketing, or even in the organization, you choose in your future career, knowledge of marketing analytics will be necessary. Marketing analytics is essential for students interested in distinguishing themselves in the job market. A study by PWC and the Business Higher Education Forum found that there were 2.35 million job postings in the U.S. for analytics-enabled jobs²¹ and the number has continued to rise. More recently, an exploration of positions listed on Burning Glass suggested an average salary of \$99,000 for analytics positions. The demand and salary for careers needing analytics training in all fields will continue to grow worldwide. Thus, job applicants with a knowledge of data science and analytics will be given preference over others without these skills.

Continued from previous page

PRACTITIONER CORNER

Theresa Kushner | Advisory Board Member at data.world and Partner at Business Data Leadership

Q How does the knowledge of marketing analytics help students differentiate themselves from peers?

A Marketing analytics is unique in that it aims the power of analytics at the customer and market—the place where companies make their money. Any student who understands how data analytics can be used to help increase revenues in a company has an edge over her peers.

Q What is the potential career impact of understanding marketing analytics and the application to business problems?

A My experience has taught me that understanding marketing analytics often grows into understanding market and business strategies. It's this understanding or experience that can catapult a career. Most companies are very interested in how they can apply analytics to their overall business as they move to transform to the digital age.

In the following chapters, marketing problems using a variety of powerful analytics software are discussed. The software platforms in Exhibit 1-12 are useful in many careers, particularly marketing, and are currently used by many organizations globally to solve complex business problems. Exhibit 1-13 displays the top software and the relative usage by companies based on a survey by KDnuggets, a leading provider of information on AI, analytics, big data, data mining, data science, and machine learning. Results were based on asking the question: What software did you use for analytics, data

Exhibit 1-12 Example of Types of Analysis Introduced

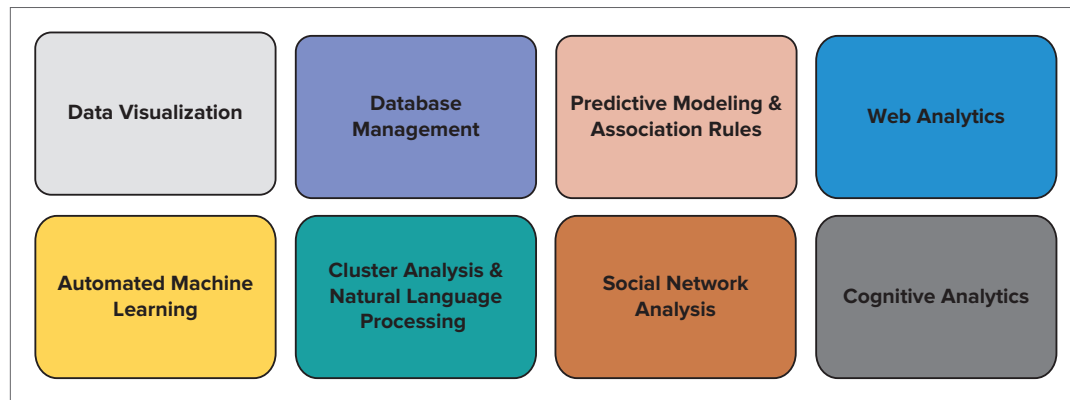
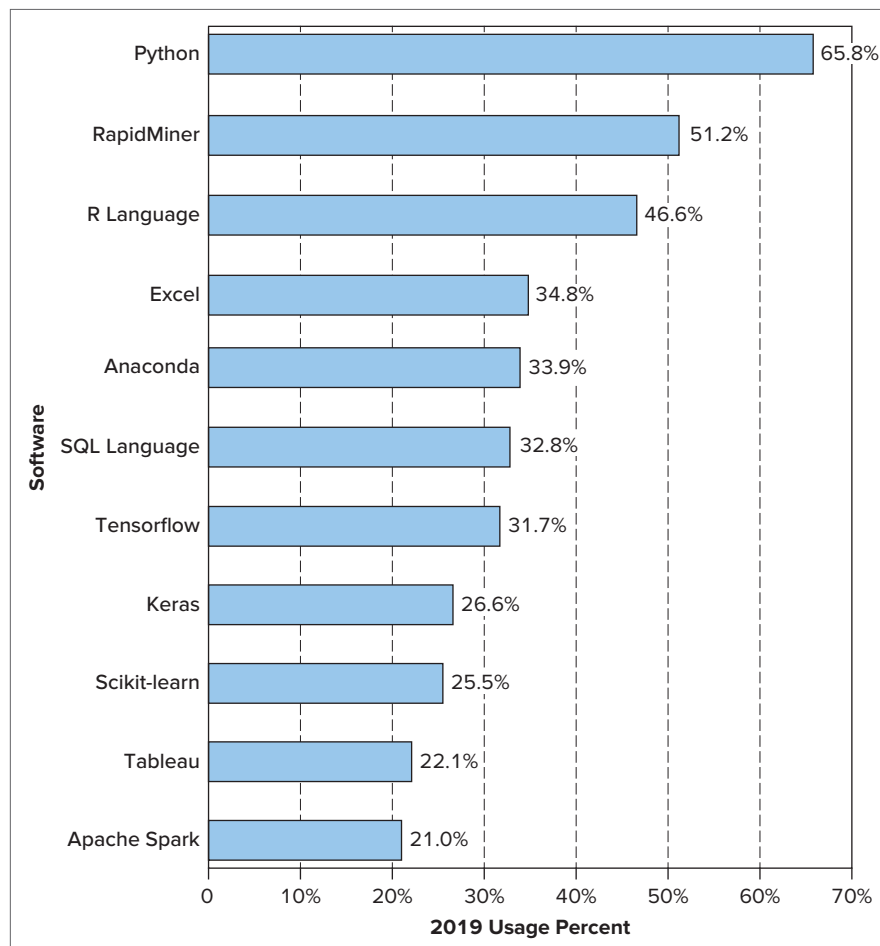


Exhibit 1-13 Top Analytics, Data Science, Machine Learning Software in KDNuggets Poll
















Source: Gregory Piatetsky, "Python Leads the 11 Top Data Science, Machine Learning Platforms: Trends and Analysis," *KDNuggets*, May 2019, <https://www.kdnuggets.com/2019/05/poll-top-data-science-machine-learning-platforms.html> (accessed June 23, 2019).

mining, data science, and machine learning projects in the past 12 months? In the remaining chapters of this book, you will learn how to solve specific marketing analytics problems through step-by-step instructions using a variety of software.

Results were based on asking the question: What software did you use for analytics, data mining, data science, and machine learning projects in the past 12 months?

This book provides the foundation and skills for successfully using analytical methods in marketing. Chapter cases cover a variety of tools and techniques to provide hands-on experiences. The variety of methods and tools you will learn in this book will offer you a toolbox approach to solving different analytics problems. Exhibit 1-14 provides an overview of the chapters by topic, modeling type, software, and coverage of the four major areas of analytics: descriptive, predictive, prescriptive, and AI/cognitive. It is time now to start your journey to develop an understanding of the fundamental marketing analytics technologies, techniques, and business applications.

Exhibit 1-14 Topics and Software Coverage in This Textbook

TITLE	TYPE OF MODELING	SOFTWARE	CHAPTER	DESCRIPTIVE	PREDICTIVE	PRESCRIPTIVE	AI/COGNITIVE
Data Management	Data Query	SQL Lite Online	Chapter 2				
AI and Cognitive Analytics	Business Intelligence Analytics	IBM Cognos	Chapter 3				
Visualization	Scatter Plot Geographic Map Heat Map Bar Chart Box Plot Network Graph Timeseries Line Graph	Tableau	Chapter 4				
Supervised Modeling	Linear Regression Neural Network Automated Machine Learning	Rapid Miner & DataRobot	Chapters 5, 6, and 7				
Unsupervised Modeling	Association Rules Cluster Analysis	Rapid Miner & Python	Chapters 8 and 9				
Natural Language Processing	Sentiment Analysis	Python	Chapter 10				
Social Network Analysis	Network Structure	Polinode	Chapter 11				
Web Analytics	Page View Click Through Engagement Time Conversion Optimization	Google Analytics	Chapter 12				

(Bulb): olegganko/Shutterstock

Summary of Learning Objectives and Key Terms

LEARNING OBJECTIVES

- Objective 1.1** Define marketing analytics.
- Objective 1.2** Discuss how to identify the right business problem.
- Objective 1.3** Identify different data sources.
- Objective 1.4** Describe different data types.
- Objective 1.5** Explain the difference between predictors and target variables.
- Objective 1.6** Differentiate between supervised and unsupervised modeling.
- Objective 1.7** Discuss the 7-step marketing analytics process.
- Objective 1.8** Ethical considerations
- Objective 1.9** Explain the value of learning marketing analytics.

KEY TERMS

Artificial intelligence (AI)	Integer	Secondary data
Binary	Interval	SMART principles
Categorical data	Machine learning	Structured data
Cognitive analytics	Marketing analytics	Supervised learning
Continuous data	Nominal	Testing dataset
Dependent, target, or (outcome) variable	Ordinal	Training dataset
Descriptive analytics	Predictive analytics	Unstructured data
Discrete data	Prescriptive analytics	Unsupervised learning
Independent variable	Primary data	Validation dataset
	Ratio	Variables

Discussion and Review Questions

1. What is marketing analytics?
2. How are companies using marketing analytics to make strategic marketing decisions?
3. Name several external data sources that might be helpful to marketers.
4. How might a company use structured and unstructured data to better understand customers?
5. Define a target variable.
6. Discuss the difference between supervised and unsupervised learning.
7. What are the steps of the marketing analytics process?
8. Discuss how companies can improve trust with their customers when building marketing applications using machine learning.

Critical Thinking and Marketing Applications

1. Visit www.data.gov. Click on Consumer, then click on Data. How many datasets are currently located on this website for free? Select one dataset and develop a scenario where the data might be helpful for a marketing

manager. Discuss how exploring the data could guide the marketing manager in making more informed decisions.

2. Develop two questions that an airline company might be interested in answering. Describe types of unstructured and structured data that might be important to answering the questions. What data sources might be helpful?

References

1. Dana Mattioli, "On Orbitz, Mac Users Steered to Pricier Hotels," *The Wall Street Journal*, August 23, 2012, <https://www.wsj.com/articles/SB10001424052702304458604577488822667325882> (accessed June 23, 2019).
2. Bernard Marr, "The Amazing Ways Spotify Uses Big Data, AI and Machine Learning to Drive Business Success," *Forbes*, October 30, 2017, <https://www.forbes.com/sites/bernardmarr/2017/10/30/the-amazing-ways-spotify-uses-big-data-ai-and-machine-learning-to-drive-business-success> (accessed June 23, 2019).
3. "Turning Fashion by the Numbers into A Billion-Dollar Business," *PYMNTS*, February 21, 2019, <https://www.pymnts.com/news/retail/2019/stitch-fix-algorithm-data-innovation> (accessed June 23, 2019).
4. Jared Council, "AI Helps Restaurant Chains Pick Sites for New Stores," *The Wall Street Journal*, May 13, 2019, <https://www.wsj.com/articles/ai-helps-restaurant-chains-pick-sites-for-new-stores-11557739802?ns=prod/accounts-wsj> (accessed June 23, 2019).
5. Kashmir Hill, "How Target Figured Out a Teen Girl Was Pregnant Before Her Father Did," *Forbes*, February 16, 2012, <https://www.forbes.com/sites/kashmirhill/2012/02/16/how-target-figured-out-a-teen-girl-was-pregnant-before-her-father-did> (accessed June 23, 2019).
6. Eric Knorr, "Hot Property: How Zillow Became the Real Estate Data Hub," *InfoWorld*, April 25, 2016, <https://www.infoworld.com/article/3060773/hot-property-how-zillow-became-the-real-estate-data-hub.html> (accessed June 23, 2019).
7. Neel Mehta, Parth Detroja, and Aditya Agashe, "Amazon Changes Prices on Its Products about Every 10 Minutes—Here's How and Why They Do It," *Business Insider*, August 10, 2018, <https://www.businessinsider.com/amazon-price-changes-2018-8> (accessed June 23, 2019); and Jia Wertz, "6 Surefire Ways to Gain Sales Traction on Amazon FBA," *Forbes*, September 28, 2018, <https://www.forbes.com/sites/jiawertz/2018/09/28/6-surefire-ways-to-gain-sales-traction-on-amazon-fba> (accessed June 23, 2019).
8. "Kellogg Realigns Supply Chain to Help Achieve Cost-Savings Goals," *MHI*, <http://s354933259.onlinehome.us/mhi-blog/kellogg-realigns-supply-chain-to-help-achieve-cost-savings-goals> (accessed June 23, 2019); Clara Lu, "Kelloggs Supply Chain Process: From Factory to Supermarket Shelves," *TradeGecko*, July 22, 2014, <https://www.tradegecko.com/blog/supply-chain-management/supply-chain-management-factory-supermarket-shelves-kelloggs> (accessed June 23, 2019); and Gurjit Degun, "Kellogg's Looks to Supply Chain to Save £300 Million," *Supply Management*, November 6, 2013, <https://www.cips.org/en/Supply-Management/News/2013/November/Kelloggs-looks-to-supply-chain-to-save-300-million> (accessed June 23, 2019).
9. Matt Marshall, "How Olay Used AI to Double Its Conversion Rate," *Venture Beat*, July 19, 2018, <https://venturebeat.com/2018/07/19/how-olay-used-ai-to-double-its-conversion-rate> (accessed June 23, 2019); and Erica Sweeney, "Olay Doubles Conversion Rates with AI-Powered Skincare Advisor," *Marketing Dive*, July 20, 2018, <https://www.marketingdive.com/news/venturebeat-olay-doubles-conversion-rates-with-ai-powered-skincare-advisor/528229> (accessed June 23, 2019).
10. J. Murawski, "Car Companies Curb AI Efforts" April 11, 2019, *The Wall Street Journal Online*, <https://www.wsj.com/articles/car-companies-curb-ai-efforts-11554888601>.
11. Bernard Marr, "The Amazing Ways Hitachi Uses Artificial Intelligence and Machine Learning," *Forbes*, June 14, 2019, <https://www.forbes.com/sites/bernardmarr/2019/06/14/the-amazing-ways-hitachi-uses-artificial-intelligence-and-machine-learning> (accessed June 23, 2019).
12. George T. Doran, "There's a S.M.A.R.T. Way to Write Management's Goals and Objectives," *Management Review* (AMA FORUM) 70(11): 35–36 (1981); and Graham

to *Leading Your Team: How to Set Goals, Measure Performance and Reward Talent* (Pearson Education, 2013), pp. 37–39.

13. Yoni Heisler, “Walmart’s Creepy Plan to Detect Unhappy Customers,” *New York Post*, July 20, 2017, <https://nypost.com/2017/07/20/walmarts-creepy-plan-to-detect-unhappy-customers> (accessed June 23, 2019).
14. Kyle L. Wiggers, “Chick-fil-A’s AI Can Spot Signs of Foodborne Illness from Social Media Posts with 78% Accuracy,” *Venture Beat*, May 23, 2019, <https://venturebeat.com/2019/05/23/chick-fil-as-ai-can-spot-signs-of-foodborne-illness-from-social-media-posts-with-78-accuracy> (accessed June 23, 2019).
15. Lauren Johnson, “Taco Bell’s Mobile Ads Are Highly Targeted to Make Users Crave Its Breakfast Menu,” *AdWeek*, March 14, 2016, <https://www.adweek.com/digital/taco-bells-mobile-ads-are-highly-targeted-make-users-crave-its-breakfast-menu-170155> (accessed June 23, 2019); Iris Dorian, “Aki Technologies Takes in \$3.75 Mln Seed,” *PE Hub*, September 2016, <https://www.pehub.com/2016/09/aki-technologies-takes-in-3-75-mln-seed/#> (accessed June 23, 2019); Geoffrey Fowler, “It’s the Middle of the Night. Do You Know Who Your iPhone Is Talking To?” *The Washington Post*, May 28, 2019, <https://www.msn.com/en-us/news/technology/its-the-middle-of-the-night-do-you-know-who-your-iphone-is-talking-to/ar-AAC1Wv1> (accessed June 23, 2019); and Lucy Sanovy, “Taco Bell Tracks Phone User Habits to Target Its Mobile Ads,” *Mobile Commerce Press*, March 17, 2016, <http://www.mobilecommercepress.com/taco-bell-tracks-phone-user-habits-target-mobile-ads/8521558> (accessed June 23, 2019).
16. “Walmart to Share Inventory Data with Suppliers in Battle with Amazon,” *Reuters*, January 30, 2018, <https://www.reuters.com/article/us-walmart-suppliers/walmart-to-share-inventory-data-with-suppliers-in-battle-with-amazon-idUSKBN1FJ1S0> (accessed June 23, 2019); Dan O’Shea, “Walmart Shares Inventory Data, Tightens Deadlines for Suppliers,” *Retail Dive*, January 30, 2018, <https://www.retaildive.com/news/walmart-shares-inventory-data-tightens-deadlines-for-suppliers/515962> (accessed June 23, 2019); and Kayla Webb, “Walmart Takes on Amazon by Sharing Inventory Data with Suppliers,” *Deli Market News*, January 30, 2018, <https://www.delimarketnews.com/retail/walmart-takes-amazon-sharing-inventory-data-suppliers/kayla-webb/tue-01302018-1130/5486> (accessed June 23, 2019).
17. Steven Melendez and Alex Pasternack, “Here Are the Data Brokers Quietly Buying and Selling Your Personal Information,” *Fast Company*, March 2, 2019, <https://www.fastcompany.com/90310803/here-are-the-data-brokers-quietly-buying-and-selling-your-personal-information> (accessed June 23, 2019).
18. Cah, “Beauty in the Age of Individualism: Sephora’s Data-Driven Approach,” *Harvard Business School*, November 13, 2018, <https://rctom.hbs.org/submission/beauty-in-the-age-of-individualism-sephoras-data-driven-approach> (accessed June 23, 2019); and K.C. Cheung, “Sephora Uses AI to Transform the Way Its Customers Shop,” *Algorithm-X Lab’s Artificial Intelligence Newsletter*, January 26, 2019, <https://algorithmxlab.com/blog/sephora-uses-ai-transform-way-customers-shop> (accessed June 23, 2019).
19. Gil Press, “Cleaning Big Data: Most Time-Consuming, Least Enjoyable Data Science Task, Survey Says,” *Forbes*, March 23, 2016, <https://www.forbes.com/sites/gilpress/2016/03/23/data-preparation-most-time-consuming-least-enjoyable-data-science-task-survey-says> (accessed June 23, 2019).
20. Casey Johnston, “Netflix Never Used its \$1 Million Algorithm due to Engineering Costs,” *Wired*, April 16, 2012, <https://www.wired.com/2012/04/netflix-prize-costs> (accessed June 23, 2019).
21. “Investing in America’s Data Science and Analytics Talent,” *PWC*, <https://www.pwc.com/us/dsa-skills> (accessed June 23, 2019).

2

Data Management

LEARNING OBJECTIVES

- 2.1** Define big data and summarize the journey from big data to smart data.
- 2.2** Discuss database management systems, relational databases, and structured query language.
- 2.3** Investigate the key elements of enterprise data architecture.
- 2.4** Define the dimensions of data quality and describe the importance of performing marketing analytics.
- 2.5** Explain the importance of understanding and preparing data prior to engaging in analytics.



Eugenio Marongiu/Image Source

2.1 The Era of Big Data Is Here

A marketing executive in a medium-sized U.S. retailer was surprised after reviewing the sales reports. One of the company's major competitors has been rapidly gaining market share. The executive was confused by the loss of market share because the firm had invested a large amount of money in improving their product design and online promotions. Upon reading a news article that examined decisions leading to the competitor's success, the executive was surprised by the challenge ahead. The competitor was investing heavily in collecting, integrating, and analyzing data from each of their stores and every sales unit. The competitor had integrated information technology (IT) infrastructure with the supplier databases, which enabled it to place orders automatically on high-demand items and shift product delivery from one store to another with ease. From e-commerce to in-store experiences, as well as across the supply chain, the competing company had become nimble and adaptive in the marketplace. What the competitor had witnessed was the game-changing impact of big data and analytics. Big data helps companies track demand and sales in real time, adapt quickly to market changes, and predict how customers will behave, thereby enabling them to personalize customer experiences.

In recent years, we have witnessed an explosion in the volume of data produced and stored. You have probably seen statements such as 90 percent of the world's data has been created in the last two years, or an estimated 2.5 quintillion bytes of data are generated every single day. However, that pace is only accelerating with data from web applications, social media, mobile apps, and sensors embedded in almost all everything we use. Large datasets are a core organizational asset that generates new opportunities and creates significant competitive advantages. Indeed, companies that adopt data-driven decision making typically achieve up to 6 percent higher productivity and output than their peers.

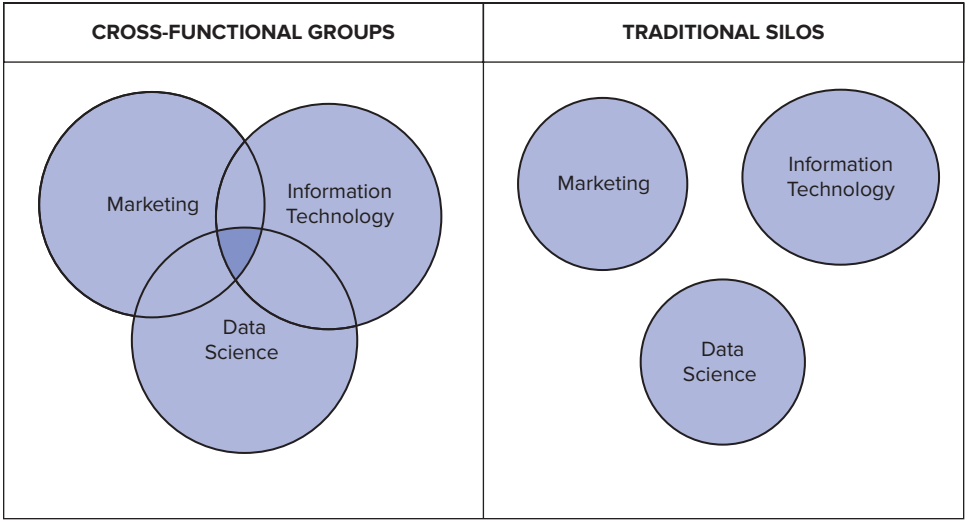
The Coca-Cola company is a good example of a business that has rebuilt itself on data to drive product development, customer retention, and engagement. With more than 500 drink brands sold in more than 200 countries, the Coca-Cola Company's customers consume more than 1.9 billion servings every day. Moreover, the company launched Cherry Sprite using data from self-service drink dispensers that allow customers to mix their own drinks. Finally, Coca-Cola identified the most popular flavor combinations and made them available to its customers.¹

Another successful data-driven company is eBay, with an estimated 179 million active buyers and more than 1.2 billion live listings across 190 markets. eBay has invested in data management infrastructure that enables it to use two decades of data and customer behavior insights to train intelligent models to anticipate the needs of buyers, recommend items, and inspire every shopper based on interest and passions.² Mastercard too has built a data infrastructure that can process 75 billion transactions per year across 45 million global locations on the Mastercard network.³ To do so, the company has moved from static data and fixed rules to fast-moving, real-time streams of transaction data, enhanced by external aggregated customer demographics and geographic information. In addition, using machine learning models, Mastercard has been able to prevent billions of dollars' worth of fraud.

Building a smart data infrastructure that enables real-time decision making requires a strong data strategy that is well aligned with overall business strategy. Many companies have adopted a data-driven strategy that encourages the use of data throughout the firm for decision making. With increased integration of both internal and external data, many companies are discovering that inter-department collaboration is a key success factor in optimizing data use. Analytics teams are

increasingly interdisciplinary, and functional departments no longer operate in silos (see Exhibit 2-1). As a result, marketers are finding themselves collaborating with numerous departments throughout the company to create successful initiatives. In addition, new departments have emerged, such as data science, that focus on statistical computational algorithms and computer programming to extract insights and knowledge from data.

Exhibit 2-1 Collaborative Workgroups versus Traditional Silos



Companies such as General Electric and Zeta Global, a marketing services company, are developing cross-functional, collaborative structures between marketing, IT, and data science to ensure more accurate data collection and identification of useful insights.⁴ Increasing applications of analytics and interdisciplinary collaboration mean marketers must have a basic understanding of data management fundamentals even if they do not have primary responsibility for managing these systems.

Integrated information technology infrastructure systems are collecting and maintaining massive amounts of data. In the airline industry, manufacturers utilize thousands of sensors and sophisticated real-time digital systems to collect more data. By 2026, over half of all wide-bodied aircraft are expected to produce 98 million terabytes of data. These advanced systems generate information surrounding engines, fuel consumption, crew deployment, and weather updates, to mention a few, with the goal of enhancing the customer experience.⁵

Multi-channel interactions are also producing large amounts of data. For example, data is generated from more than 100 million desktop and mobile visits to the Target website each month, 40 million users that downloaded the mobile app, as well as by customers shopping at over 1,800 brick and mortar stores.⁶ Other types of external data are also collected, such as social media mentions, weather patterns, and economic developments. Another retailer, Walmart, processes over 2.5 petabytes (PB) of internal and external data every hour, seven days a week. But realistically, is it possible for us to understand how much data this truly is? To put this into perspective, a petabyte (PB) is a million gigabytes (GB), and a single petabyte can store about 225,000 movies. Imagine looking pixel by pixel at your favorite 225,000 movies, and that is only one petabyte. Consider the most recent thumb drive or cloud storage service you used. What was the storage capacity? While they range in size, many thumb-drive storage options for personal use hold 100 or more gigabytes (GB), and external laptop drives

PRACTITIONER CORNER

Peter Drewes | Business Intelligence, Intellectual Property and Strategy
at Lockheed Martin



Peter Drewes

Dr. Peter Drewes is a Business Intelligence and Intellectual property manager at Lockheed Martin focusing on the sustainment areas of the F35 program. His background is idea valuation and the business, strategic, and technical coordination necessary to bring those to life. Over his 30+ year career, he has developed and helped launch multi-billion dollar opportunities in unmanned systems, autonomous applications, and supercomputing. This has involved entering new markets for Underwater autonomous vehicles, transforming geospatial data into collective team knowledge, and advancing robotics through research into data analysis. His business focus has been that of combining business analytics and market analysis into cohesive strategic plans.

Q Peter, from your perspective why is it critical for employees working in any function to understand or participate in the company's processes for data management, from collection to preparation, analysis, and ultimately strategy development?

A The goal for every function is to understand the lifecycle of the data and processes that are being generated. What questions need to be answered? What requirements are placed on the data and the analysis? How long will the data and information be relevant, and how do you know data biases have been reduced so the information will be useful in answering questions?

The key is for those who ask questions to understand data lifecycle elements so they can ask and get answers to the right questions. For example, they must be able to bridge the silos between functional departments to provide detailed answers. Typically, in this situation, the marketing department is the question

generator that wants to understand how customers are making decisions. As part of the journey, the marketing team will therefore have to bridge the gaps between silos to answer the fundamental questions. If marketers are uninvolved in the process or unable to collaborate, other departments such as data scientists must have the same understanding as the marketing department, but they seldom do. Functional departments can be adept in many areas, but the data science department does not make strategic marketing decisions. Similarly, the IT department does not specify how data scientists do their analysis. It is this bridging of the gaps between groups that facilitates appropriate technology investments, achieves desired returns on investment goals, while at the same time meeting their strategic goals. Utilizing each department's expertise is the only effective way to reach individual department and overall company goals.

Continued to page 37

hold a terabyte (TB) or more. For an understanding of computer technology measurement storage units, refer to Exhibit 2-2.


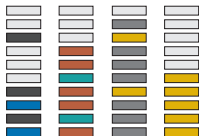
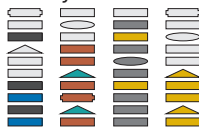
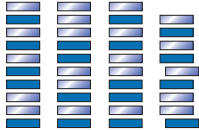
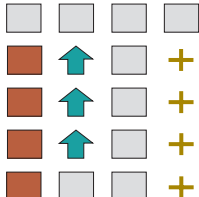
Big data is the term typically used to describe massive amounts of data. It is a relative term that requires special tools to manage and analyze. The amount of data collected by Target and Walmart would be considered big data by most standards. But Amazon, Google, and Facebook are other companies that have amassed large amounts of customer search history and purchase data. The existence of big data is a result of the digital transformation taking place in companies and among customers. Companies not only store historical and real-time data in a digital format, they also interact with suppliers and customers using a variety of digital methods that contributes to big data.

Exhibit 2-2 Computer Technology Storage Units of Measure

UNIT OF MEASURE	NUMBER OF BYTES	EXAMPLES ⁷
Yottabyte (YB)	1000 ⁸	as much information as there are atoms in 7,000 human bodies
Zettabyte (ZB)	1000 ⁷	as much information as there are grains of sand on all the world's beaches
Exabyte (EB)	1000 ⁶	about one-fifth of the words people have ever spoken
Petabyte (PB)	1000 ⁵	half of the contents of all U.S. academic research libraries
Terabyte (TB)	1000 ⁴	all the X-rays in a large hospital
Gigabyte (GB)	1000 ³	approximately 500 high-quality audio songs
Megabyte (MB)	1000 ²	a small novel
Kilobyte (KB)	1000	a paragraph of a text document
Byte (B)	1	a single character of text

The word big, however, is somewhat subjective. In the case of big data, several characteristics illustrate the term (see Exhibit 2-3). Volume, variety, veracity, velocity, and value are several characteristics used to describe big data.

Exhibit 2-3 Characteristics of Big Data

Volume	Large Data at Rest 	Companies must now store and analyze petabytes of data. Data is collected from a variety of sources and enables companies to examine the entire customer journey.
Variety	Diverse Data 	Data can range from structured to unstructured. There are strengths and challenges to these different formats when creating an integrated database, but variety provides a more holistic understanding of customers and market situations.
Veracity	Messy Data 	The data could have missing values, inconsistencies in the unit of measurement, erroneous information, and lack of reliability, which increases complexity and reduces confidence in the data.
Velocity	Fast Data in Motion 	Troves of data are being produced by digital technology. This data is inundating companies at a rapid pace (taking milliseconds to seconds to send). This speed supports real-time response strategies.
Value	Useful Data 	The extracted data must be converted into quality insights that add tangible and intangible benefits to the business. Achieving value requires an understanding of the goals and objectives of the business.

Volume refers to large amounts of data per time unit. The volume of data can be anticipated in the case of regular purchase behavior, but intense public attention on social media can bring inconsistent volumes of data, an amount a company might be unprepared to manage. Companies such as Walmart report more than 140 million customers visit a typical company-owned brick-and-mortar or ecommerce site each week.⁸ These customers produce a vast quantity of data through purchase transactions, returns, browsing patterns, and search history. As you can imagine, the large quantity of data arrives with high **velocity** or speed. This high-volume, high-speed data moves constantly between the network of exchange relationships from suppliers to retailer stores to customers. Managing the volume and speed of incoming data can be challenging due to the potential **veracity** or **variety** of data. Recall from Chapter 1 that structured and unstructured types of data are constantly generated from various sources. Making a purchase online or in a store would yield structured data such as names, addresses, phone numbers, and purchase amounts. In contrast, data originating from social media through videos, text, and images would likely be unstructured. It is easy to understand how collecting similar data from several sources increases the potential for inconsistencies in the units of measure or missing data.

Data is available to almost all companies throughout the world, but it is only an asset when it provides **value**. Value means the data is useful for making accurate decisions. Many professionals are moving away from the term “big data” and beginning to adopt the term “smart data.” **Smart data** represents data that is valuable and can be effectively used. Big data should be well-organized and made smart prior to analysis by making sure it can be used to produce more accurate decisions.

2.2 Database Management Systems (DBMS)

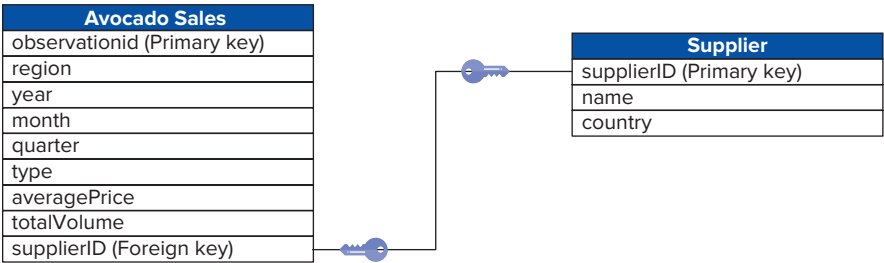
Have you ever considered how big data is organized to create smart data that provides value? All data is stored and organized in a database. A **database** contains data collected from company operations. The data must be organized for efficient retrieval and analysis by different functional departments throughout the company. When employees search for customer information or customers search for products online, a database is working behind the scenes to provide the best results. Customer relationship management (CRM) and product search systems are commonly stored in relational databases.

A **relational database** is a type of database management system (DBMS). It is a collection of interrelated data items organized using software programs to manipulate and access data. The software programs involve a framework, often referred to as a schema, that captures the structure of how the database is constructed with tables, columns, data types, stored procedures, relationships, primary keys, foreign keys, validation rules, etc. A relational database stores data in structured columns and rows similar to an excel spreadsheet. Exhibit 2-4 shows sales for a specific plant variety: the Hass avocado. The relational table includes information about avocado sales and consists of a set of product or company attributes, including such items as region, average sales, total volume, and type. A set of tables is one component of a relational database, each of which has a unique name. As shown in Exhibit 2-4, a table typically includes a set of columns (also known as features, predictors or variables) and stores a large number of records. The row (also known as records) are often identified by a unique primary key and described by columns. A foreign key is a set of one or more columns in a table that refers to the primary key in another table. Primary keys and foreign keys are important in relational databases, because they help database users combine data from different tables, as shown in Exhibit 2-5.

Exhibit 2-4 Example of a Table Structure

OBSERVATION ID	REGION	AVERAGE PRICE	TOTAL VOLUME	TYPE
1	Albany	1.47	113514.4	Organic
2	Atlanta	0.95	649352.6	Conventional
3	Baltimore/ Washington	1.15	849487.6	Conventional
4	Boise	1.13	79646.97	Conventional
5	Boston	1.4	419696.6	Organic
6	Buffalo/Rochester	1.27	115508.3	Organic

Exhibit 2-5 Example of a Relational Database Structure



How can data in relational databases be accessed for greater meaning? Relational data is accessible by a database management language called **structured querying language (SQL)**. The language was developed by IBM and is used to access and update data stored in the database. A query can be used to join, select, manipulate, retrieve, and analyze data from relational databases. These databases are beneficial when data consistently maintain the same properties because they require predefined structures. If the company decides to begin collecting customer email addresses or locational information, the database tables would need to be altered to accept any new column.

On the other hand, non-relational databases (see Exhibit 2-6), also known as NoSQL databases, can store large volumes of structured or unstructured data. Non-relational databases show data vertically, combined together rather than in structured tables. For example, the first row in Exhibit 2-6 matches the first column of Exhibit 2-4 and refers to organic Hass avocados in Albany.

Exhibit 2-6 Non-Relational (NoSQL) Database Example

```
{“Name”: “Average Sales Price: 1.47”},
{“Name”: “Total Volume: 113514.4”},
{“Name”: “Type: organic”},
{“Name”: “Region: Albany”}
```

NoSQL databases allow greater flexibility for storing ever-changing data and new data types, but drilling down to very specific types of data is more difficult. The flexibility of NoSQL databases is important for companies with dynamic sources of data, such as mobile devices or social media.

Most companies use both relational and non-relational type databases to store data. Data often resides in multiple sources with different data formats. As expected, managing data efficiently can be challenging. The difficulty of maintaining multiple databases is compounded by inappropriate data storage architecture.

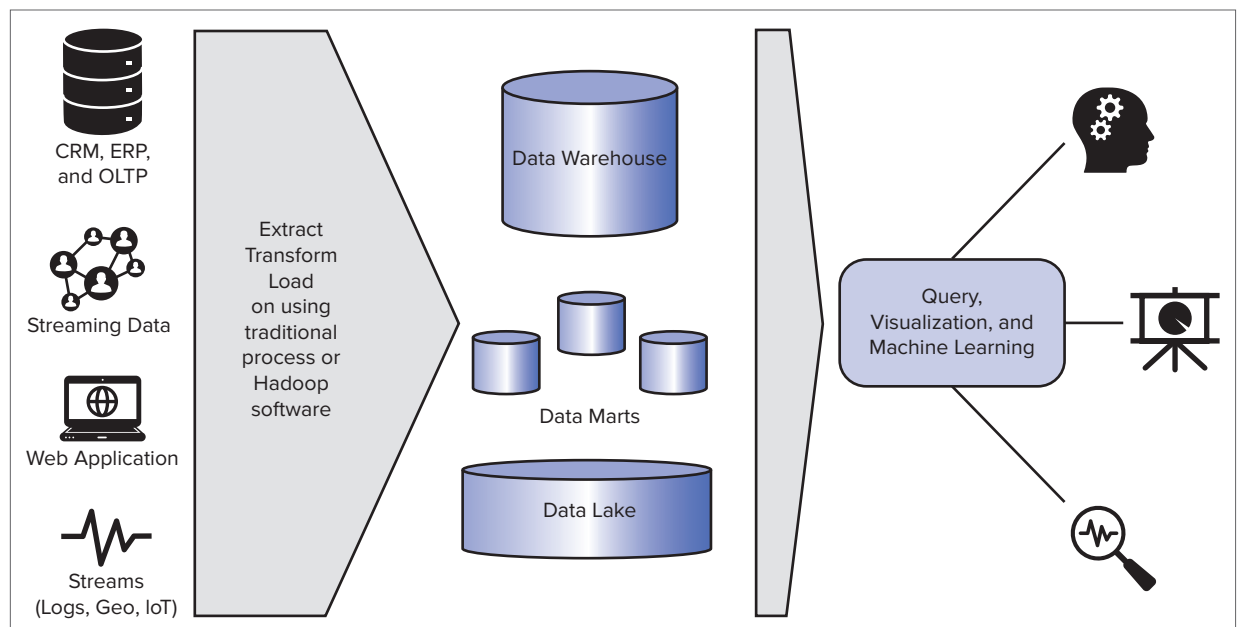
2.3 Enterprise Data Architecture

Data storage architecture provides a framework for companies to systematically organize, understand, and use their data to make both small and large decisions. In recent years, companies have spent millions of dollars building enterprise-wide data architecture to help drive informed decision making in a fast-changing world. Many organizations view their data architecture as a competitive position to help them retain customers by learning more about their needs. Interestingly, new trends have evolved in data architecture with more companies investing in flexible cloud infrastructure and open-source architecture.

Exhibit 2-7 shows the basic architecture of a data storage environment for an organization. In principle, data analytics can be applied to analyze any kind of information repository. This includes Customer Relationship Management (CRM), Enterprise Resource Planning (ERP), and other Online Transaction Processing (OLTP) software that supports operational data transactions such as customer order entry, financial processing, material order, shipping process, and customer service. As an example, CRM solutions use an operational database to store customer data. Some of the most common CRM solutions are available through Microsoft, SAP, Salesforce, and Oracle. A CRM database might store recent customer transactions or responses to marketing promotions and allow marketers to monitor developments in real time. The data storage environment offers a place to combine different sources of data. Internal company data can be combined with data from other sources such as social media (e.g., Instagram, YouTube, Facebook) to capture customer opinions about products, reviews, and video comments, which also can be combined with web sales data.

Streaming data, the continuous transfer of data from numerous sources in different formats, can also be included to capture customer data. This might include geographic mobile data, sensor data from physical stores, and logs from internal systems and web capturing of the type, content, and time of transactions made by the user interacting with the system.

Exhibit 2-7 Simple Architecture of a Data Repository



Traditional ETL

Extract, Transform, and Load (ETL) is an integration process designed to consolidate data from a variety of sources into a single location (see Exhibit 2-8). The functions begin with *extracting* key data from the source and converting it into the appropriate format. For example, the date is converted into data/time format. A rich transformation process then includes cleaning the data, applying transformation, and name conversion. *Transformation* requires conforming to the appropriate data storage format for where data will be stored. For example, a CRM database might require structured columns and rows such as first names, last names, and telephone numbers. But text from email communications, customer service interactions, or social media would be considered unstructured and must be specified differently. Because both structured and unstructured data are important, ETL solutions are being improved to efficiently integrate these various types of data. Most traditional ETL tools can process only relational datasets for semi-structured, unstructured data, and machinery sensor data, but newer systems are much more flexible. The third ETL step is *load*, in which the data is loaded into a storage system such as data warehouse, data marts, or a data lake.

Exhibit 2-8 Functions of Extract, Transform, Load (ETL)

FUNCTIONS
Data is extracted from the source.
Data is transformed into a useable form.
Data is integrated and loaded into a storage system.

ETL Using Hadoop

The massive volume of data led to the development of new technologies like Hadoop to capture, store, process, secure, and then analyze complex data (Exhibit 2-7). **Hadoop** is an open-source software that helps distributed computers solve problems of big data computation. Hadoop divides the big data processing over multiple computers, allowing it to handle massive amounts of data simultaneously at a reduced cost. Hadoop also facilitates analysis using MapReduce programming. MapReduce is a programming platform used to manage two steps with the data. The first step is to map the data by dividing it into manageable subsets and distributing it to a group of networked computers for storing and processing. The second step is to combine the answers from the computer nodes into one answer for the original problem handled. HIVE, a data warehouse built using Hadoop, provides SQL-like query to access data stored in different file systems and databases that are used by Hadoop.

The loading process uses an open-source Hadoop framework, reducing the cost of operation. Most importantly, the ETL process on Hadoop can handle structured, semi-structured, and unstructured data. After the ETL process is completed using traditional ETL or Hadoop, data can be stored in a data warehouse, data marts, or a data lake.

A Closer Look at Data Storage

One popular database architecture, a **data warehouse**, contains historical data from various databases throughout a company and provides a structured environment for high-speed querying. A data warehouse consists of data from different functional areas of the firm and likely includes data associated with customers, human resources, and accounting. The data is typically organized under one schema to facilitate holistic decision making in the organization. Merck, a healthcare company, recently faced a data problem. Employees were spending as much as 80 percent of their work time gathering data. The result was not enough time to complete

developed a data warehousing system where data scientists can analyze both structured and unstructured data at the same time, and also develop reports using data visualization software for business analysts. Even though the databases might be located in different departments, the data warehouse system provides a central repository.

A **data mart** is a subset of the data warehouse that provides a specific value to a group of users. For example, a marketing data mart would be limited to data on customers, sales, products, and similar marketing metrics. There are two types of data marts: One is referred to as a dependent data mart, in which the data is directly obtained from an enterprise data warehouse. The second is referred to as independent data mart, in which the data is obtained from transactional systems, external providers, or specific geographic areas.

A **data lake** (often included in Hadoop systems) is a storage repository that holds a large amount of data in its native format. It is typically used to investigate data patterns and to archive data for future use. A variety of big data sources such as social media, weather data, logs, and sensor data, as well as online reviews, semi-structured (HTML) and unstructured data (e.g., videos, pictures), and machine-generated data (e.g., sensor data) can all be stored in a data lake. With data lakes, millions of customer records can be explored quickly without having to wait for data to be loaded into the data warehouse. At the same time, the contents of data lakes can be integrated with data warehouse. TD Bank Group is the sixth-largest bank in North America, employing around 85,000 people. TD Bank Group recently began to transform its digital infrastructure to a data lake architecture hosting customers' personal data such as demographics, preferences, opinions, and other external structured and unstructured data using a Hadoop private cloud.¹⁰ Using tools like Hive, Apache Spark, and Tableau, the bank was able to explore massive amounts of data and build insightful reports and visualizations in a very cost-effective and quick manner. The data architecture enabled the company to move from "data-silos" to democratized access, providing information to employees across the organization. This data democratization enabled TD Bank to offer adaptive and customized products and services to its customers.

Consider the value of an enterprise-wide data repository for multi-channel retailers. Multi-channel retailers collect operational, transactional, and social media data from various sources. Customers make rapid decisions based upon information available to them at the right time. Companies rely on the real-time integration of multi-channel data to facilitate customer decision making through price modifications or streamlining inventory that help maintain customer satisfaction.

Continued from page 31

PRACTITIONER CORNER

Peter Drewes | Business Intelligence, Intellectual Property and Strategy
at Lockheed Martin

Q **How are companies managing data from a variety of sources to create a full view of data and generate consistent value?**

A With modern computing power and inexpensive (temporary) data storage, large amounts of data can be stored and analyzed at will. The fundamental question is what to do with that data 2, 3, or 4 months after it has

been stored? Is all of the data useful for market segments and analysis, or have they chosen to collect everything to prevent missing data that might be needed in the future? The key is the upfront setup of the inbound data management and quality parameters. Each department within a

(Continued)

company must be vocal and engaged in establishing a plan for data collection and management that provides answers to the questions relevant to them. It is very important to architect a solution up front to decide where the data is coming from and how often, how it will be stored, what analytics will be done on the data, and for how long. This will determine how to follow the data from field to useful information and how long you need to keep it handy.

Q How do marketers play an important role in data management?

A Marketers, business development teams, and process optimization teams work together to look at the problems and decide on the appropriate questions to answer. This will drive the requirements of data analytics that will ultimately deliver the insights to answer those questions.

The rate of internet-enabled devices/processes continues to explode. More data is available from internal and external processes. It wasn't that many years ago that 1 MB of RAM was enough to solve many problems. Today's chasing of shiny objects might include "Let's use Artificial Intelligence/Machine Learning to make sense of our large datasets." But using a highly tuned algorithm to find relationships without a specific understanding could render unreliable results. Since the current trend is to accept what the computer has generated as a valid output, it is important the process of data management include functions across the enterprise.

Q How is the process of data management evolving to facilitate tangible results across the enterprise?

A It is an exciting time when internal questions can be asked, supporting data is identified and tied to external financial market analysis, and where predictive models answer the "what if" questions: "Should we go to market?" "Should we attempt to gain marketspace against company XYZ?" "Would this method be successful?" We are no longer limited by the quantity of data available, but by the ability to frame that data into successful business decisions. It is toward this point where the process of data management is evolving.

Business transformations based on governance and processes should be supported by massive amounts of "clean" data. Just because we have the resources to invest in a particular product does not mean we should. It might take too long to market to customers, be served by alternate means, be unrecoverable in terms of realizing an investment. No longer do we make the judgment "build it and they will come" or perform similar "guessing in the dark" exercises about a potential market analysis or a social media campaign. Using appropriate data management and numerical techniques, tangible results can be realized through insights that facilitate the analysis of markets or provide a value chain assessment with guidance for future endeavors. This type of power will influence businesses for the foreseeable future. Those that can take advantage of it and manage the incoming data will succeed; those that cannot will never realize their full potential.

Continued to page 40

For the data to be useful in addressing business problems, it must be properly managed. **Data management** as a process is the lifecycle management of data from acquisition to disposal. Prior to advancements in computing, data management consisted of paper copies placed in filing cabinets. If someone needed a file, they could review the paper copy and hopefully remember to replace it in the correct location. Today, data management provides a stable solution that facilitates efficient data access to appropriate people that reside both internally, such as the marketing team, and externally, such as suppliers, to the company. The management of data includes the strategy, structure, and processes to access and store data throughout the organization. Company data often consists of diverse data sources. The marketing, sales, and customer service departments collect different types of data from different sources. Marketing might collect information pertaining to customer responses to promotions, the sales department likely maintains data on current and churning customer accounts, and the customer service department acquires data through customer interactions from telephone calls, email, or chatbots. Although the data collected is usually required in making quick decisions for the particular department, it is not held in a vault

only for their viewing. Most companies share access to data across functions or with external partners as necessary. Good data management provides a foundation for the delivery of high-quality data.

A major challenge of today's data management is that the inbound data continues to increase exponentially (Walmart, for example). These mountains of data are created each day and must be cleaned, verified, and validated so they can be used at some time in the future as high-quality data.

2.4 Data Quality

High-quality data is critical. Numerous success stories inspire current event articles recognizing the use of data by companies in decision making—to improve products, enhance customer relationships, adjust pricing strategies, and so forth. Unfortunately, not all companies experience the same level of success, because data too often contains errors that are not fixed. Coca-Cola used market research data of 200,000 people to create the New Coke product, one of the worst recorded product flops of all time. Based upon the research data, it was anticipated that consumers would adopt the new product. Unexpectedly, however, the new product development was a total failure that left customers upset, with 400,000 people expressing their dissatisfaction through phone calls and letters. The data was of poor quality and incomplete. It excluded important data, such as other predictors beyond taste—for instance, purchase behavior.

There is a common adage that people use when referring to deficient **data quality**: “garbage in, garbage out.” The statement is reflected in the Coca-Cola example. If the database contains poor quality data, results or decisions emanating from that data will also be of poor quality. We can again follow the trail of bad data through a company to recognize the impact on multiple groups of people. Inaccuracies in a supplier's inventory management system will lead to confusion for corporate retail buyers whose job it is to maintain appropriate levels of products in stores. An absence in products on the shelves leads to frustrated store managers and disappointed customers who, in today's digitally connected environment, could effortlessly search for products elsewhere. Many companies have adopted site-to-store pickup options, but if their inventory is not current, then customers could place orders expecting to receive the item within a few hours, but receive a notice indicating the product is no longer in stock. The customer must then start from ground zero by either searching and purchasing the product from a competitor or online. Inaccurate data, missing fields, or data isolated in disparate sources can also be drivers of underperforming employees and dissatisfied customers.

If you consider the growth of data sources and vast amounts of data being generated, it is easy to comprehend how the quality of data might be negatively impacted. Customers or employees can easily input incorrect data by typing an incorrect ZIP code, using an invalid email address, misspelling a name, or inputting a decimal in the wrong location. These mistakes might seem harmless, but contemplate the effect of misspelled names. Jonh Smith, Jon Smith, and John Smith happen to be the same person but are currently registered as separate accounts in the company's CRM system, resulting in disconnected historical interactions by the same customer. This becomes an issue when the company decides to identify high-value customers based on purchase history. This customer would likely be eliminated because the transactions are divided into several customer names and subsequently smaller purchase amounts.

Let's examine another scenario. Consider purchasing your car at a dealership, then returning for regular service. If you purchase a service contract, for example, your car service is free for the first two years of purchase or 25,000 miles—whichever comes first. You schedule maintenance for your car to discover that the dealership's records indicate you already had your 25,000-mile service, even though you are positive the last

maintenance visit was for 20,000 miles. They are using multiple software systems to store customer data and must investigate the claim prior to scheduling an appointment. As a result, the situation requires more of your time to search for vehicle maintenance records and to call back to verify they located the correct information. Data quality issues such as these—whether a result of incorrect data, missing data, or disconnected systems—are all too common and frequently lead to frustrated customers. Lack of quality data can also result in consequences to company employees because critical decisions are based upon data that is available to them in the system.

Continued from page 38

PRACTITIONER CORNER

Peter Drewes | Business Intelligence, Intellectual Property and Strategy
at Lockheed Martin

Q “Garbage in, garbage out.” What are some major challenges companies encounter with data quality issues?

A First and foremost is to determine how well a company follows their data governance policy. This will set the stage for everything following data collection. How is data collected, verified, and validated? How is it cleaned and parsed for later usage? These are all key elements in the reduction of the “garbage in” problem. Combining this with asking the right questions will reduce, but not eliminate, the garbage out.

A couple of common data quality issues are related to outliers or missing data. Are the outliers more distant from the curve true outliers or the data of interest? The dataset may have completely clean data that is relevant and useful, but still provides garbage output. In the case of outliers, they can be useful depending on the questions being asked, or they could produce erroneous results. Another area to consider is how missing data is addressed. One philosophy is to replace it with the mean to reduce deviations. However, there are always other questions to consider. Why is the data missing? Was there a valid reason that caused the data to be “missing”? Answers to these and similar relevant questions provide the basis to ensure usable information is provided throughout the process.

Q What is the impact of bad data on decision making?

A The major problem with bad, inconsistent, or biased data is that often times nothing jumps out to indicate the results are invalid. Technology will run analytics models for many

hours through many terabytes of data to develop conclusions, but that does not mean the results are always correct. Analysts need to use training datasets to develop initial models that identify errors and revise as necessary. But in the real world, training datasets are not applied often enough because they reduce system efficiency and increase overhead costs to get to the answer. It is this recursive training using multiple datasets, however, that allows algorithms and data to be continuously analyzed and improved. If a test set of the highest 20 percent of data is used, does that change the results? What about using the lowest 20 percent? Which datasets will change the system conclusions? Examining questions such as these will yield more valuable information. Since we are relying on systems to provide a “faster” look at the data than we can do by hand, it becomes more challenging to watch the algorithms do their job given the volume of inbound data. This level of carelessness leads to business decisions we would never make if we had better data.

Another key area is the garbage in, hallelujah out concept. When biased or bad data isn’t corrected, the results are skewed or potentially bad. However, since the algorithms and process have been approved and validated, we tend to trust the algorithms and the output data is believed to be true. Results are then often run up and down the management chain as good news (the hallelujah portion). But later when performance isn’t met, or the biases are found, the decision-making process is questioned, reducing the chance management will accept the next analytical solution.

Continued to page 42

Unfortunately, when data are of poor quality, insights produced by marketing analytics will be unreliable. Data are important, but high-quality data is critical in developing an accurate understanding of trends and purchase patterns and in maintaining customer satisfaction. For data records to be valuable, they have to be timely, accurate, complete, and consistent. Many times, unsatisfied customers are a result of poor data quality.

Consider the restaurant chain example from Chapter 1. Restaurants are using a variety of data sources and machine learning to determine where to locate their next establishment. Included in the analysis might be internal information such as existing stores and competitor locations, store sales, area demographics, and traffic patterns, and could conceivably include lifestyle information such as healthy eating preferences of area customers from social media. What might be the financial repercussions of inaccurate data from a single one of the sources mentioned? Restaurants are relying upon the combination of this data as the foundation for AI driven analyses. Results will guide decisions to optimize placement and predict sales of new locations. Inaccuracies will taint the data and produce false information. Data has the potential to be a valuable resource for companies. But poor quality data can have a significant, negative impact.

Although data quality can be measured by numerous dimensions, the most common are timeliness, completeness, accuracy, consistency, and format:

- *Timeliness*: Have you ever found a product you wanted to quickly purchase at a brick-and-mortar store, but could not access a mobile phone or internet service to determine if it was available elsewhere at a lower price? You then left the store after purchasing the product only to realize a competitor was selling it for much less. The lack of timely information led to a decision that might have changed if the information was available when needed. Similar scenarios occur in other purchasing situations. Real-time data such as customer social media sentiment, responses to marketing campaigns, online and offline customer behavior patterns, customer service chatbots, or call centers are all critical in making current decisions. If customers communicate their dissatisfaction on social media or through call centers, companies prefer to know this sooner rather than later. They can then respond as quickly as possible in an effort to lessen damage to the brand. Having a complete set of data is also important to responding in a timely manner.
- *Completeness*: Imagine if a company wanted to send an email follow-up to their best customers and did not collect email addresses or wanted to personalize a communication using a first name, but it only had access to initials. Data completeness means maintaining the sufficient breadth, depth, and scope for the particular task. In this case, the company would need to forgo or delay the email campaign or personalized communications because the data was incomplete.
- *Accuracy*: Data accuracy is the extent to which the data is error-free. Is the data correct, reliable, and accurately measured? If it is, then decisions more often result in a favorable outcome.
- *Consistency*: Data inconsistencies can lead to embarrassing dilemmas and uncoordinated strategic decision making. Consider Nestlé USA as an example. The labeling of vanilla was different from one division and factory to another. When the data was being evaluated for integration, the inconsistent values created confusion and there was no efficient way to reconcile the differences. Because the data was inconsistent, the company was unaware the same supplier was charging different prices for each division and factory, when everyone should have realistically been paying the same price.
- *Format*: Format is the extent to which the information is adequately presented or delivered for efficient and effective understanding. When one person recently visited

a Dallas, Texas, emergency room, the patient reported returning from a trip outside the country to the nurse, who documented the information. But the hospital later reported that although the records were accessible to the physician, his immediate computer screen did not contain the information and the patient was ultimately discharged. Unfortunately, it was later discovered the patient had Ebola, a highly contagious, often fatal disease and a panic ensued throughout the country. While this is a serious example, many functional departments make decisions from information dashboards. If the dashboard is difficult to read or clumsy to navigate, the impact could be detrimental.

2.5 Data Understanding, Preparation, and Transformation

For the curious mind, the idea of exploring data can be somewhat of a game. In a game, you use available information in hopes of making the best decisions, while also obtaining a high score or defeating your competitors. The data is used to satisfy an outcome—in this particular scenario, the outcome is winning. Data inspires curiosity for marketers because they are eager to use insights to strategize the next move. Marketers desire to explore data as quickly as possible because data is basically useless until analyzed. Consider rows and columns of data—there is no tangible benefit to data in raw form. But do not get distracted by the excitement of discovering fascinating new insights. There are important steps to consider in the process. Data is messy and must be tidied up before being used.

Data Understanding

Recall the 7-step marketing analytics process that was introduced in Chapter 1. First, marketers must grasp an understanding of the business. If there is a failure to understand the business situation, the marketer will have a difficult time defining the right business problem and asking the right questions.

Continued from page 40

PRACTITIONER CORNER

Peter Drewes | Business Intelligence, Intellectual Property and Strategy at Lockheed Martin

Q Why is it so important that marketing analysts have a thorough understanding of the business and existing data?

A Everything is related to asking the right question. If you ask the wrong question, you are heading toward a useless business answer. This is never done intentionally, but preventing this misstep requires a thorough understanding of the business processes, existing data, needed data, marketing, and business outcomes. If only one person understands all of this, either their

judgment must be completely trusted, or others should be brought up to speed to discuss different parts of the analysis. Without a collaborative, top-down view, everyone is operating in a silo environment. This may achieve local goals while creating system-wide inefficiencies. Thus, management and analysts must be able to see eye to eye and understand the same problem from the same perspective. System-wide efficiency can only be obtained through this understanding.

Continued to next page

Understanding available data is also critical to correctly addressing the desired business problems and reducing the potential for inaccurately reporting results. It might seem obvious to confirm you understand the data, but individual data fields are easily overlooked when dealing with large datasets. Pretend your company is struggling to understand sales over the last three years and your supervisor requests an analysis of sales data. Your job is to simply compare sales on an annual basis. You notice that the unit of measure for sales figures is reported monthly, and so you simply create an annual column for each year. In doing so, however, you do not realize the dataset only contained six months for the third year. Unfortunately, the mistake is not found until presenting the information, which erroneously reports third year sales have plummeted. Without a proper data understanding, you made an error in analyzing the data and the annual sales comparison is incorrectly reported.

Continued from previous page

PRACTITIONER CORNER

Peter Drewes | Business Intelligence, Intellectual Property and Strategy
at Lockheed Martin

Q **“Chaos to Clarity.” What are several key steps in the role of data cleaning and preparation (e.g., merging, aggregating) that are necessary for a marketing analyst to understand prior to data analysis?**

A Data analysis is the logistical execution of the data algorithms selected based on the questions to be answered. Following appropriate data management, the real work is done during the preparation and cleaning based on the question requirements. The design of the experimentation and the methods that will be used to test the data and algorithms should align with how the data is structured. Therefore, merging sources or

aggregating values might be necessary in answering certain questions.

The data governance process begins with studying what structured and unstructured data will arrive, how often it will arrive, what transport mechanism is involved (IoT, web socket, Excel sheet, or napkins). What does the timing mean compared to the decision I need to make? Having perfect data organized and correlated is a wonderful thing. But if the data is available six months after a corporate strategic decision is needed, it offers no value and can potentially hurt the company making decisions with only the data available at that moment.

Data Preparation

Extensive use of descriptive statistics and data visualization methods is necessary to gain an understanding of the data. First, the data must be properly prepared. During the process of preparing the data, tasks may include identifying and treating missing values, correcting errors and removing outliers in the data, deciding on the right unit of analysis, and determining how to represent the variables.

Feature Selection In most data analytics projects, the analyst will have access to ample data to analyze a model. For this reason, it is important to pay close attention to the variables (also known as **features** or predictors) included in the model. In situations like these, the data is likely to have a large number of features. Some of these features (e.g., person’s height and weight) might be highly correlated or measure the same thing. Other features could be completely unrelated to the variable of interest or the target variable. If features are correlated with each other or unrelated to the target variable, this can lead to poor reliability, accuracy, and what is referred to as **overfitting** of the model. But what is **overfitting**? Consider a dataset with 100 individuals, 40 percent of

whom have purchased products at the same company before. Information about interest, income, number of children, neighborhood ZIP code, and past purchase behavior might do a good job predicting whether or not someone will purchase the product. But if we keep adding additional predictors such as height, number of pets, weight, and hair color, the accuracy of the model will likely improve.

This performance may be misleading, however, because it probably includes spurious effects that are specific to the 100 individuals, but not beyond the sample.

Overfitting occurs from an overly complex model where the results are limited to the data being used and are not generalizable—which means future relationships cannot be inferred, and results will be inconsistent when using other data. Always keeping every variable available in the model is not necessary and often misleading. The analyst can determine which variables are sufficient to retain for the analysis. It is important, therefore, to know the features to include and which to eliminate to confirm that the model represents key variables of interest. Keeping too many variables can be unnecessary to achieve good model results and is costly to manage in terms of computer processing power. Typically, this step involves consulting stakeholders with domain specific knowledge (i.e., individuals with knowledge of the data and the business process) to ensure the features included are important from the end-user perspective.

Sample Size As with feature selection, **sample size** is an important consideration. Political polls often report results of a certain sample size. Surveying every voter is impossible; therefore, these analyses focus on smaller, more reasonable proportions of the population. Sample size recommendations are based on statistical analysis concepts such as a power calculation. A **power calculation** helps determine that outcomes will be estimated from a sample with a sufficient level of precision. In a data mining project with a large dataset, the goal typically is not to estimate the *effect size* but to predict new records. When the goal is accurate prediction, a large sample of several thousand records is often required.

Unit of Analysis A **unit of analysis** describes the what, when, and who of the analysis. To identify the appropriate unit of analysis, the first step is to identify the target (outcome) variable. Let's say we are trying to predict whether a customer is likely to churn (leave our business), then the customer is the unit of analysis. On the other hand, if the purpose is to predict brand satisfaction, then the brand is the unit of analysis. If a business is trying to determine the price of a mobile phone subscription plan for different customer levels, then the unit of analysis is the subscription rather than the person. At times, the unit of analysis may not be obvious. Therefore, it is a good practice to collaborate with other subject matter experts to collectively determine the correct unit of analysis.

Missing Values It is common to have missing records for one or more of the variables in the dataset. Missing values result from various scenarios such as customers that, due to time constraints, overlook and leave fields blank. This creates a problem because analysts must make adjustments for missing data. There are several options to address missing data: (1) Imputing missing values with estimated new values (mean, median, regression value), (2) omitting the records with missing values, and (3) excluding the variables with the missing values.

If among 35 variables, the average sale is missing for only three records in that variable, we might substitute the mean sale for the missing records. This will enable us to proceed with the analysis and not lose information the record has for the other 34 complete variables. The imputation options are recommended when the missing values are missing at random and typically include the mean, median, and mode. For example, suppose the average sale for all customers is \$45,000, marketing analysts could decide

to use this value to replace the missing value for sales. Another option with missing values is to predict the missing value using regression or decision tree induction. For example, customer sales values might be missing. Relying on the other customer attributes in a dataset, we could use a decision tree to predict the missing values for sales. The missing values can then be replaced with predicted values using the results of the decision tree.

Removing incomplete observations when the number of missing values is small is also a reasonable option. It is important, however, to consider if the missing values are missing at random or if there is a pattern or reason for the missing values. For example, in a study of the relationship between product satisfaction and product performance, if participants with an above-average satisfaction skip the question “Rank the product performance,” analyses may not identify the relationship between product performance and satisfaction. Information such as this is essential to understanding the relationships between variables.

When observations are missing for a large number of records, dropping the records will lead to a substantial loss in the dataset. It is good in cases like this to examine the importance of the variables with large missing values. If the variable is insignificant in the prediction of the model, it can easily be excluded. If the variable is a key indicator, however, then marketers must determine if it is worth the investment to collect data and obtain the missing records. When working with large datasets, removing many observations with missing values might have no effect on the overall analysis. This is particularly valid when the variable dropped has a high correlation with another variable in the dataset, and thus the loss of information may be minimal.

Outliers Values that are at a considerable distance from any of the other data clusters are considered **outliers**. In large datasets, outliers are typically detected using a statistical test that assumes a distribution model of the data or distance measures. As a rule of thumb, an outlier is often determined as anything over three standard deviations from the mean. Although not grounds for immediate removal, outliers should be investigated further. For example, we know that an age observation of 140 indicates an error. However, it might be determined that an observation of 100 is within the possibility of accurate data and should be retained for the analysis. In cases when only a few outliers exist, removing them from the dataset is warranted. Most of the time, outliers need to be removed from the model as noise. In some applications such as fraud detection, however, these rare outlier events are what the model is trying to predict.

Consider several customer income levels in a dataset: \$5,000, \$45,000, \$48,000, \$50,000, \$1,000,000. As a marketer, you are developing pricing strategies based upon customer demographics. In this simple case, two values stand out as obvious. The income levels of \$5,000 and \$1,000,000 do not conform with the general behavior of the data. You would need to decide whether the customers with an income of \$5,000 or \$1,000,000 would skew the results of purchase behavior based on the price of your product. To examine existing outliers in a data analytics project, we can review the maximum and minimum values for each variable. Are there values falling toward the minimum or maximum points of the value distribution? Here, the customer reporting an income of \$1,000,000 would fall toward the maximum point of distribution. Another method of identifying outliers is to use **cluster analysis**, where groups are created based upon similarities to determine if any observations have a considerable distance to other clusters.

For a more detailed explanation of how to assess and deal with these data issues, refer to Hair, Babin, Anderson, and Black (2018).¹¹ Once data has been cleaned, the marketing analyst should determine whether the data warrants further transformation through aggregation, normalization, new variable construction, or dummy coding.

Data Transformation

- **Aggregation:** During this process, summary operations are applied to the data. For example, the weekly sales data may be aggregated to calculate the total sales by month, quarter, or year. This is a key process to help prepare data at the unit of analysis necessary for insights.
- **Normalization:** Variables (also known as features or predictors) may include different data ranges that vary greatly from each other. Some data might be measured weekly and other data annually. To normalize a variable, we scale it by subtracting the variable from the mean and then dividing it by the standard deviation. Normalization helps us bring all variables into the same scale. Normalization becomes critical in some data analytics techniques, such as cluster analysis, because large values will dominate the distance calculation.
- **New column (feature) construction:** A new feature (predictor or variable) can be calculated based on other existing values. If a dataset consists of a sales date, a marketer might want to know more about whether sales are more prevalent on certain days of the week, months, or seasonally. Using the sales date, new columns of data can be constructed by the day of the week, month, quarter, and year.
- **Dummy Coding:** This process can be useful when considering nominal **categorical variables**. A categorical variable is when the data represents one of a limited number of categories. Geographic location (e.g., Northeast, Southeast, Northwest, Southwest, Midwest) is an example of a categorical variable. In this case, geographic location is considered a nonmetric variable and need to be re-coded using a process called dummy coding. Dummy coding involves creating a dichotomous value from a categorical value. This type of coding makes categorical variables dichotomous using ones and zeroes. Dummy coding is covered in more detail in Chapter 5.

Case Study Avocado Toast: A Recipe to Learn SQL

One of Americans' favorite "go to" breakfasts is avocado toast. In fact, it is reported that Americans spend almost \$900,000 on avocado toast each month. How has this growth impacted avocado sales over the last few years? We can use SQL to query basic results that answer questions such as this and more. In this section, we will take a closer look at avocado sales data over four years across the United States. In exploring this data, you will learn how to manipulate thousands of rows using SQL language. This exercise will provide you with the knowledge to SELECT tables in SQL, JOIN data, CREATE and INSERT new tables, and modify data using UPDATE.

Getting Started

As mentioned earlier in this chapter, one of the most popular databases is a relational database. It is a framework made of tables (similar to a spreadsheet) where each row (also known as a record) represents a set of related data describing a customer or a product. On the other hand, each column (also known as a feature, predictor, or variable) represents data such as company name, region, street address, quantity sold, and so on. Exhibit 2-9 shows a subset of data in the avocado dataset. Columns include region, average price, and type. The rows contain the data for this table (e.g., the first row shows Phoenix, 0.77, conventional).

Exhibit 2-9 A Subset of the Avocado Data (Year = 2018)

REGION	AVERAGE PRICE	TYPE
Phoenix	0.77	Conventional
Charlotte	1.23	Conventional
Denver	1.52	Organic
San Diego	1.15	Conventional

As previously mentioned, the way to manipulate data in a database is by using a query language. One of the most popular query languages is SQL. To demonstrate how SQL can be used to add and modify data, we will use SQLITE, a web version of SQL. SQLite is an introductory tool, but it will enable us to explore the basics of SQL programming. When you are ready to build a full application using an SQL server, this basic knowledge will help you get started.

Understanding the Dataset

This data can be downloaded from the student's resources page. Data for this exercise was downloaded from the Hass Avocado Board website but modified for this case study.¹² The data represents compiled weekly avocado sales for more than three years and comes from retailer's cash registers based on retail sales of Hass avocados. The Average Price (of avocados) in the table reflects a per-unit (per avocado) cost, even when multiple units (avocados) are sold in bags. To get started, let's review the data elements in the table (Exhibit 2-10).

Exhibit 2-10 Data Elements Represented in the Hass Avocado Data

VARIABLE NAME (TYPE)	DESCRIPTION
observationid (typeless)	A unique identifier for each observation. This is a <i>primary key</i> that is located across multiple avocado datasets. A primary key can guide the integration of data from one table to the data of another table.
region (string)	The sales geographic location
year (date)	The year of the observation
month (date)	The month of the observation
quarter (date)	The quarter of the observation
type (string)	Conventional or organic
averageprice (numeric)	The average price of a single avocado
totalvolume (numeric)	Total number of avocados sold
supplierid (typeless)	An identifier indicating the supplier of the avocado. Note that supplierid is a <i>foreign key</i> in the avocado table.

There are several tasks you will learn in SQL via the case study on how to prepare data for analysis and answer questions about it.

After reviewing the data, what questions can you ask to better understand avocado sales since 2016? Here are some ideas:

- What are the highest average prices customers are paying for a single avocado?
- What is the average price customers are paying per region?
- Where are avocados being sold over certain average prices?
- What is the average price that customers are paying in specific geographic regions?
- How can the data be aggregated to obtain the volume of conventional avocados per quarter versus per week?
- How are regions performing by volume sales?
- What is the company and country of origin for avocados supplied to each region?

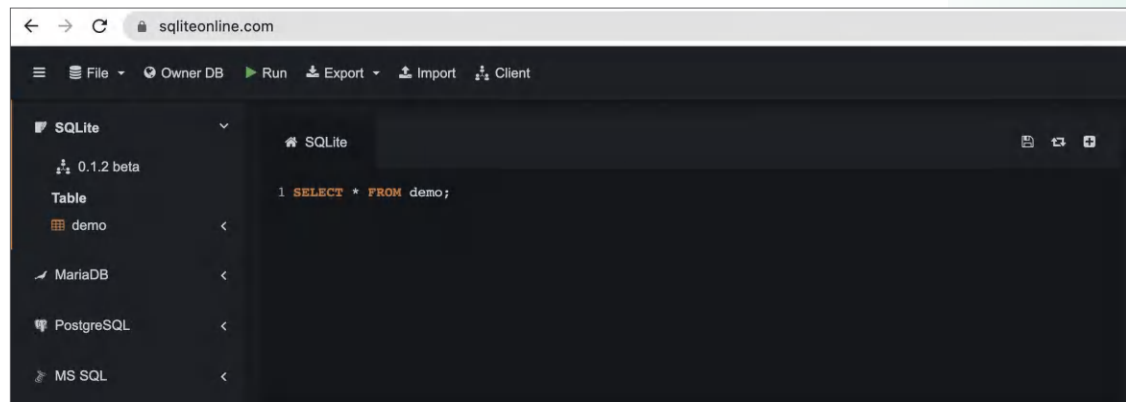
You can use SQL for basic data analysis to answer these questions and more.

Applying the Concepts

There are several options for exploring SQL. For this short introductory exercise, you will use an online platform.

Step 1: Visit the website <https://sqliteonline.com>. Your screen should look like Exhibit 2-11.

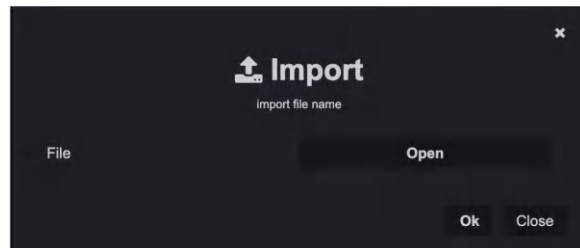
Exhibit 2-11



SQLite

Step 2: To explore the data, it must be imported into the SQLite Online platform (Exhibit 2-12). Click on “Import” and then “Open.”

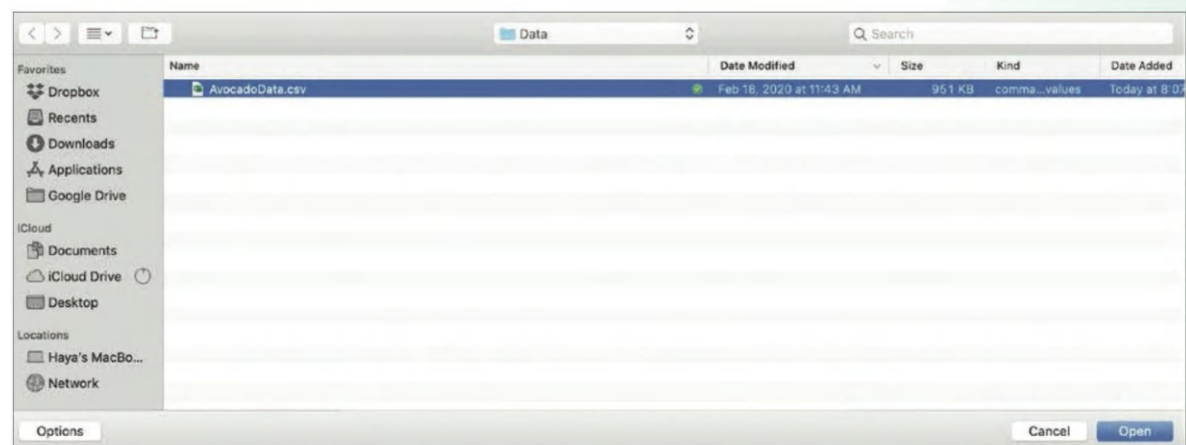
Exhibit 2-12



SQLite

Step 3: Click “Open” and browse for the csv file “AvocadoData” that you downloaded from the (need location for files) page (Exhibit 2-13).

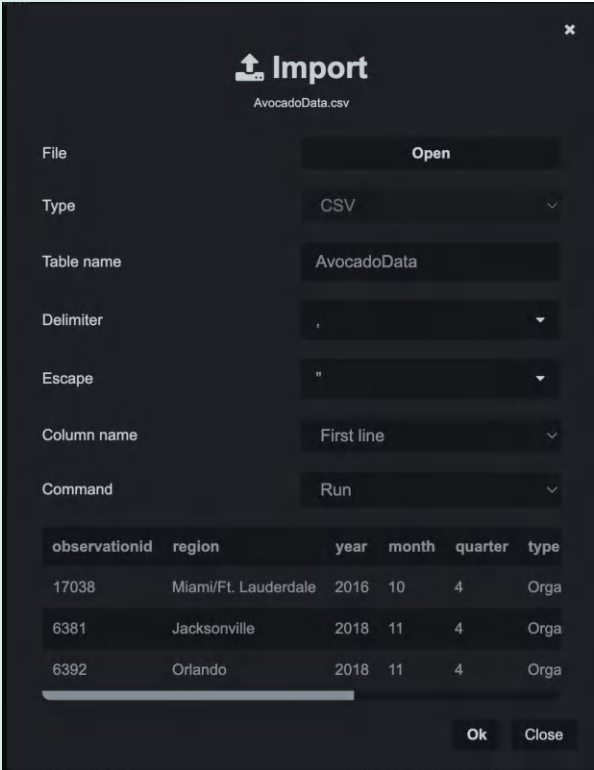
Exhibit 2-13



SQLite

Step 4: After selecting “Open,” the data import specification will appear. Update “Column name” from the drop-down menu to “First line.” Your selected options should match Exhibit 2-14.

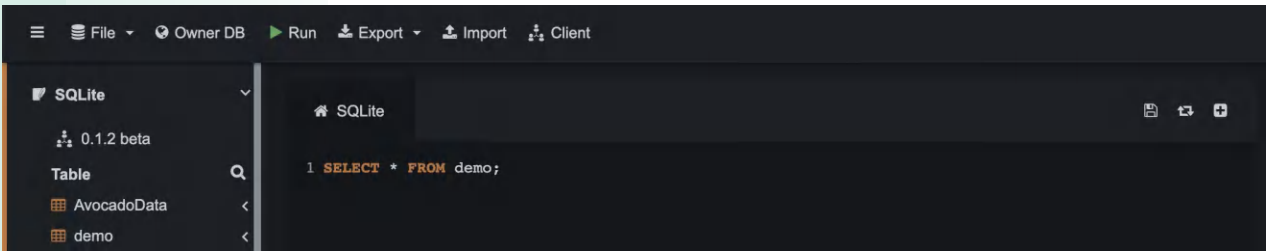
Exhibit 2-14



SQLite

Step 5: After selecting “Ok,” the data will upload and your screen should appear like that shown in Exhibit 2-15.

Exhibit 2-15

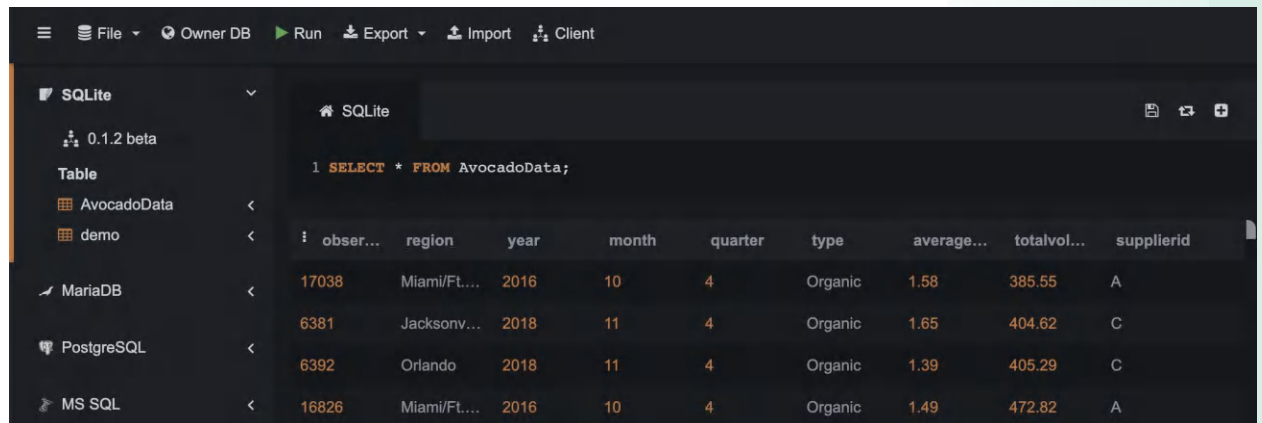


SQLite

Step 6: Now, let’s preview the data. Perform this task by using the following query statement and clicking “Run” (Exhibit 2-16):

```
SELECT * FROM AvocadoData;
```

Exhibit 2-16



The screenshot shows a database client interface with a sidebar on the left listing databases: SQLite (0.1.2 beta), AvocadoData, demo, MariaDB, PostgreSQL, and MS SQL. The main window displays the SQLite database with the query `1 SELECT * FROM AvocadoData;` entered. Below the query, a table of results is shown with columns: obser..., region, year, month, quarter, type, average..., totalvol..., and supplierid. The data rows are:

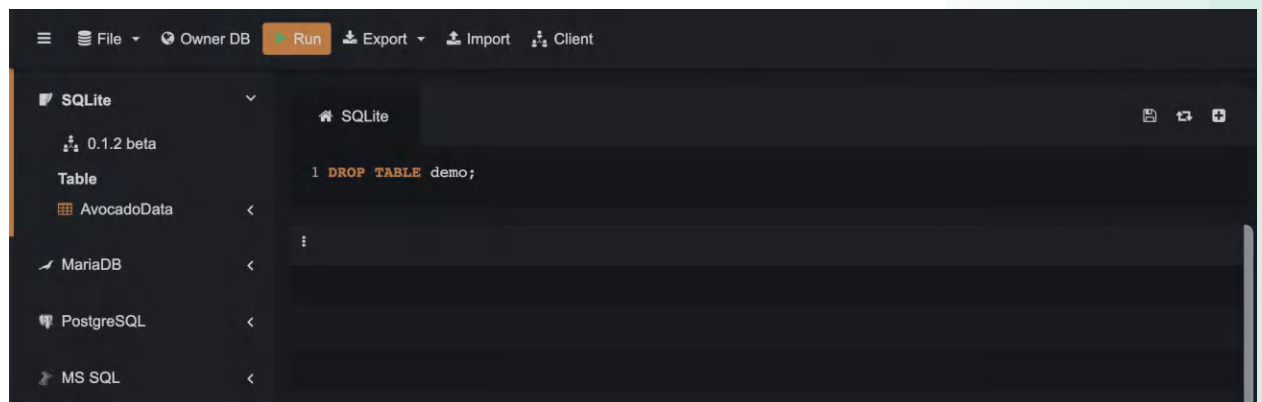
observed	region	year	month	quarter	type	averageprice	totalvolume	supplierid
17038	Miami/Ft...	2016	10	4	Organic	1.58	385.55	A
6381	Jacksonv...	2018	11	4	Organic	1.65	404.62	C
6392	Orlando	2018	11	4	Organic	1.39	405.29	C
16826	Miami/Ft...	2016	10	4	Organic	1.49	472.82	A

SQLite

Step 7: In this step, you will learn how to drop (or delete) a table from the database. Perform this task by using the following query statement and clicking “Run” (Exhibit 2-17):

Drop TABLE demo;

Exhibit 2-17



The screenshot shows the same database client interface as Exhibit 2-16, but the query entered in the SQL editor is `1 DROP TABLE demo;`. The sidebar on the left remains the same, listing the databases.

SQLite

Step 8: You can view selected fields and sort your data. Suppose you are interested in just looking at the average price paid for an avocado by region but want to sort results by average price from high to low. Perform this task by using the following query statement and clicking “Run” (Exhibit 2-18):

```
SELECT averageprice, region, year FROM AvocadoData ORDER BY averageprice DESC;
```

From sorting the data from high to low, we see that San Francisco has the highest average price from 2016.

SQLite

1 SELECT averageprice, region, year FROM AvocadoData ORDER BY averageprice DESC;

averageprice	region	year
3.25	San Francisco	2016
3.17	Tampa	2017
3.12	San Francisco	2016
3.05	Miami/Ft. Lauderdale	2017
3.04	Raleigh/Greensboro	2017
3.03	Las Vegas	2016
3	Portland	2017
3	San Francisco	2017
2.99	Indianapolis	2017
2.99	San Francisco	2016
2.97	Raleigh/Greensboro	2017
2.96	Seattle	2017
2.95	Southeast	2017
2.94	Southeast	2017
2.94	San Francisco	2016
2.93	Spokane	2017

SQLite

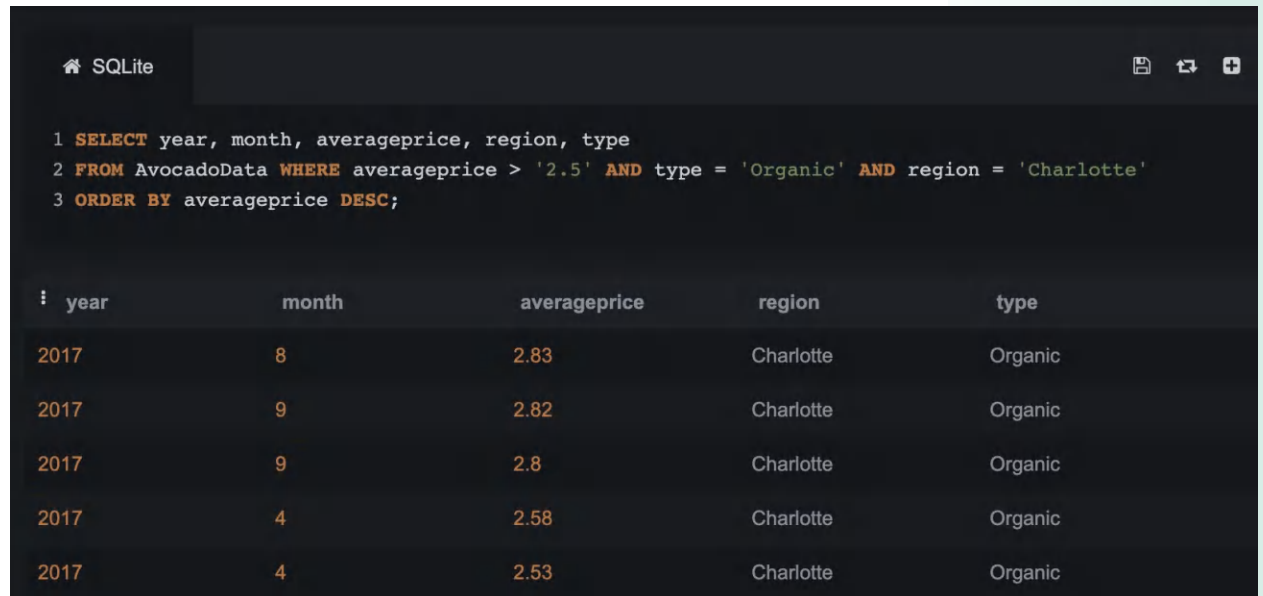
Step 9: What if you want to focus on reviewing a smaller subset of the data? To create smaller subsets, use a conditional statement (WHERE) to meet a condition. For example, you would like to examine sales where the average price is a certain value. You can include additional criteria (e.g., region) to further narrow your query. To include this type of criterion (region), use what is referred to as a logical statement (AND, OR).

Maybe you would like to examine the average price of organic avocados in Charlotte, North Carolina, that are greater than \$2.50. In this query, you will need to include some conditions that would limit the returned data specifically to the defined geographic area. Perform this task by using the following query statement and clicking “Run” (Exhibit 2-19):

```
SELECT year, month, averageprice, region, type
FROM AvocadoData WHERE averageprice > '2.5' AND type = 'Organic' AND region =
'Charlotte'
ORDER BY averageprice DESC;
```

As you can see from the results, the highest average price of organic avocados sold in Charlotte, North Carolina, was \$2.83 in August 2017.

Exhibit 2-19



The screenshot shows the SQLite application interface. At the top, there is a toolbar with icons for home, SQLite, save, share, and a plus sign. Below the toolbar, the SQL query is entered in a text area:

```
1 SELECT year, month, averageprice, region, type
2 FROM AvocadoData WHERE averageprice > '2.5' AND type = 'Organic' AND region = 'Charlotte'
3 ORDER BY averageprice DESC;
```

Below the query, the results are displayed in a table with the following columns: year, month, averageprice, region, and type. The results are sorted by averageprice in descending order.

year	month	averageprice	region	type
2017	8	2.83	Charlotte	Organic
2017	9	2.82	Charlotte	Organic
2017	9	2.8	Charlotte	Organic
2017	4	2.58	Charlotte	Organic
2017	4	2.53	Charlotte	Organic

SQLite

Step 10: To better understand the total volume of conventional avocados in quarter 1 of the last three years, sort by year from high to low. Perform this task by using the following query statement and clicking “Run” (Exhibit 2-20):

```
SELECT region, Round((totalvolume),0) as totalvolume, averageprice, month, year
FROM AvocadoData WHERE quarter= '1' AND type = 'Conventional' ORDER BY
year DESC;
```

Results indicate that, in the first quarter of 2019, a total of 79,041 conventional avocados were sold in the Syracuse region at an average price of \$1.16 in January.

SQLite

1 SELECT region, Round((totalvolume),0) AS totalvolume, averageprice, month, year FROM AvocadoData

2 WHERE quarter = '1' AND type = 'Conventional' ORDER BY year DESC;

region	totalvolume	averageprice	month	year
Syracuse	79041	1.16	1	2019
Syracuse	79294	1.19	2	2019
Spokane	82073	1.3	2	2019
Syracuse	85397	1.18	3	2019
Syracuse	85815	1.2	3	2019
Syracuse	88470	1.1	2	2019
Albany	89104	1.25	1	2019
Syracuse	90713	1.12	1	2019
Syracuse	92371	1.11	2	2019
Syracuse	94651	1.11	1	2019
Spokane	95753	1.18	3	2019
Boise	100176	1.09	2	2019
Boise	100904	1.16	3	2019
Spokane	101850	1.2	1	2019
Spokane	103262	1.21	1	2019


SQLite




Step 11: Another way to return a selected set of data that meets multiple criteria is using the statement IN. For example, if you want to understand only two cities from the results in step 6, consider returning all the fields to limit the focus on just those two cities—in this case, Boston and San Francisco. Perform this task by using the following query statement and clicking “Run” (Exhibit 2-21):

```
SELECT * FROM AvocadoData WHERE region IN ('Boston', 'San Francisco');
```

Now, only results from Boston and San Francisco are visible.




Exhibit 2-21

 SQLite

```
1 SELECT * FROM AvocadoData WHERE region IN ('Boston', 'San Francisco');
```

id	observed_on	region	year	month	quarter	type	averageprice	totalvolume	supplierid
16435		Boston	2016	8	3	Organic	1.23	7117.91	A
16329		Boston	2016	7	3	Organic	1.14	7313.26	A
14739		Boston	2016	1	1	Organic	1.26	7629.12	A
11188		Boston	2017	9	3	Organic	1.47	7698.45	C
14792		Boston	2016	1	1	Organic	1.32	7751.94	A
16965		Boston	2016	10	4	Organic	1.32	8153.98	A
14951		Boston	2016	1	1	Organic	1.52	8221.86	A
10696		San Francisco	2017	7	3	Organic	2.45	8311.12	C
16541		Boston	2016	8	3	Organic	1.36	8395.34	A
17018		Boston	2016	10	4	Organic	1.49	8408	A
16912		Boston	2016	10	4	Organic	1.43	8760.34	A
14898		Boston	2016	1	1	Organic	1.46	8850.23	A
16950		San Francisco	2016	10	4	Organic	2.34	9048.18	A
14845		Boston	2016	1	1	Organic	1.13	9327.19	A
15110		Boston	2016	2	1	Organic	1.52	9394.21	A
15004		Boston	2016	2	1	Organic	1.62	9462	

SQLite

Aggregation

Most times when you get a dataset, you will need to roll it up to a higher level. For example, you can calculate the maximum average avocado price in our dataset, or the sum of total Volume by quarter or year. To do this, it will be necessary to add a function to the variable you would like to roll up. For numeric type variables (e.g., averageprice, totalvolume), you can use the structured query language (SQL) aggregate functions for a set of values: sum, min, max, average, and so on (Exhibit 2-22). For categorical data (e.g., region and type), you can use functions such as count. When using aggregate functions, the result will be produced on a single row.

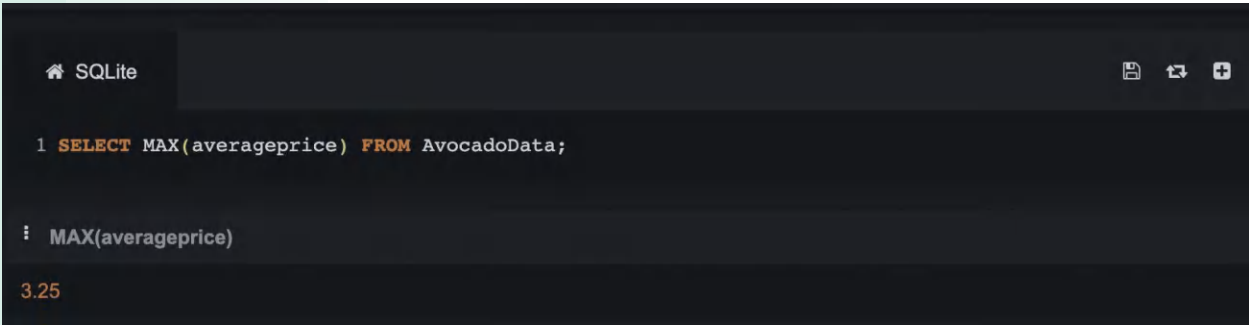
Exhibit 2-22 Aggregate Functions

FUNCTION	RETURNS
AVG ()	The average value of the selected group.
COUNT ()	The number of rows that correspond to a certain feature.
MAX ()	The maximum value in a group.
MIN ()	The minimum value in a group.
SUM ()	The sum of values within a group.

Step 12: The averageprice column contains the average price of avocados by city and date. What if you want to search for the highest average price that customers have paid to date? Perform this task by using the following query statement and clicking “Run” (Exhibit 2-23):

SELECT MAX(averageprice) FROM AvocadoData;

Exhibit 2-23



SQLite

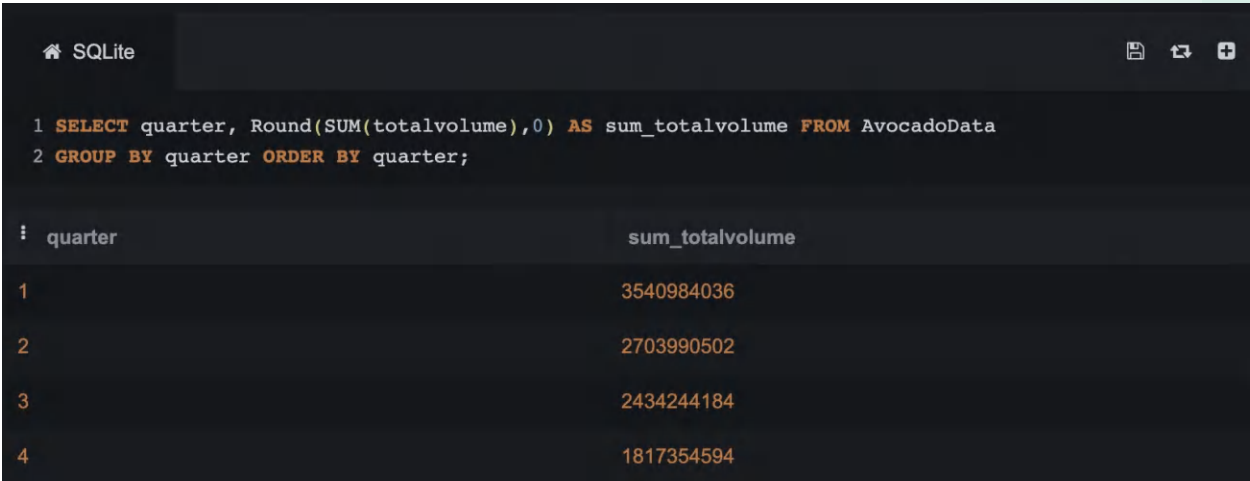
Results indicate that the highest average price that customers paid was \$3.25. You can also use the MIN(averageprice) statement to find the lowest average price that customers paid to date. What is the lowest price that customers paid to date?

Step 13: Suppose your data contains the month/day/year in a single column, but you would like to examine data over each quarter instead. You can aggregate the total sales volume by summing the values by quarter. Perform this task by using the following query statement and clicking “Run” (Exhibit 2-24):

SELECT quarter, Round(SUM(totalvolume),0) as sum_totalvolume FROM AvocadoData GROUP BY quarter ORDER BY quarter;

Using the Round function allows us to specify the number of decimal places for totalvolume. In this example, the number of decimals is set to zero. As you can see during quarter 1, customers from all regions purchased over 3.5 billion avocados. Now you can see the total avocado purchases by customers from all regions for each quarter.

Exhibit 2-24



The screenshot shows the SQLite application window. At the top, there's a toolbar with icons for file operations. Below it, a text area contains the following SQL query:

```
1 SELECT quarter, Round(SUM(totalvolume),0) AS sum_totalvolume FROM AvocadoData
2 GROUP BY quarter ORDER BY quarter;
```

Below the query, the results are displayed in a table with two columns: 'quarter' and 'sum_totalvolume'.

quarter	sum_totalvolume
1	3540984036
2	2703990502
3	2434244184
4	1817354594

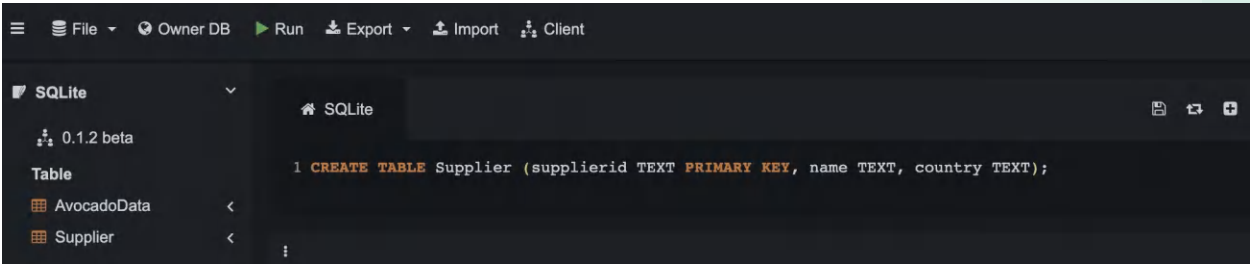
SQLite

Build Your Own Supplier Table

Step 14: The dataset also contains supplier information. If you need to know which supplier is providing avocados to certain cities, you might want to develop another table with the supplier information. Perform this task by using the following query statement and clicking “Run” (Exhibit 2-25):

```
CREATE TABLE Supplier (supplierid TEXT PRIMARY KEY, name TEXT, country TEXT);
```

Exhibit 2-25



The screenshot shows the SQLite application window with a sidebar on the left. The sidebar has a 'Table' section with two entries: 'AvocadoData' and 'Supplier'. The main text area contains the following SQL query:

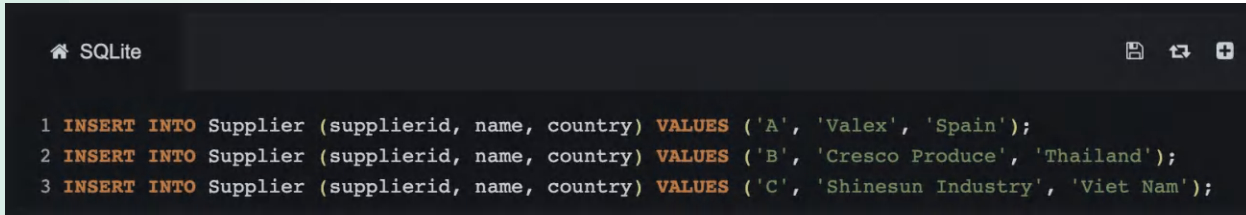
```
1 CREATE TABLE Supplier (supplierid TEXT PRIMARY KEY, name TEXT, country TEXT);
```

SQLite

Add Data to Your Table

Step 15: To add data to your data, you can insert supplier id, name, and country location. Perform this task by using the following query statement and clicking “Run” (Exhibit 2-26):

```
INSERT INTO Supplier (supplierid, name, country) VALUES ('A', 'Valex', 'Spain');
INSERT INTO Supplier (supplierid, name, country) VALUES ('B', 'Cresco Produce', 'Thailand');
INSERT INTO Supplier (supplierid, name, country) VALUES ('C', 'Shinesun Industry', 'Viet Nam');
```



```

SQLite
1 INSERT INTO Supplier (supplierid, name, country) VALUES ('A', 'Valex', 'Spain');
2 INSERT INTO Supplier (supplierid, name, country) VALUES ('B', 'Cresco Produce', 'Thailand');
3 INSERT INTO Supplier (supplierid, name, country) VALUES ('C', 'Shinesun Industry', 'Viet Nam');

```

SQLite

Join the Two Tables (MERGE)

Step 16: Currently, you have two tables. Table Avocado consists of observationid, region, year, month, quarter, type, averageprice, totalvolume, and supplierid, and table Supplier includes supplierid, name, and country. Note that supplierid is a common key between the Avocado table and the Supplier table. It is a primary key in the Supplier table, and it is a foreign key in the Avocado table. Thus, we can select attributes from both tables by joining the tables using the common key (supplierid). Perform this task by using the following query statement (Exhibit 2-27):

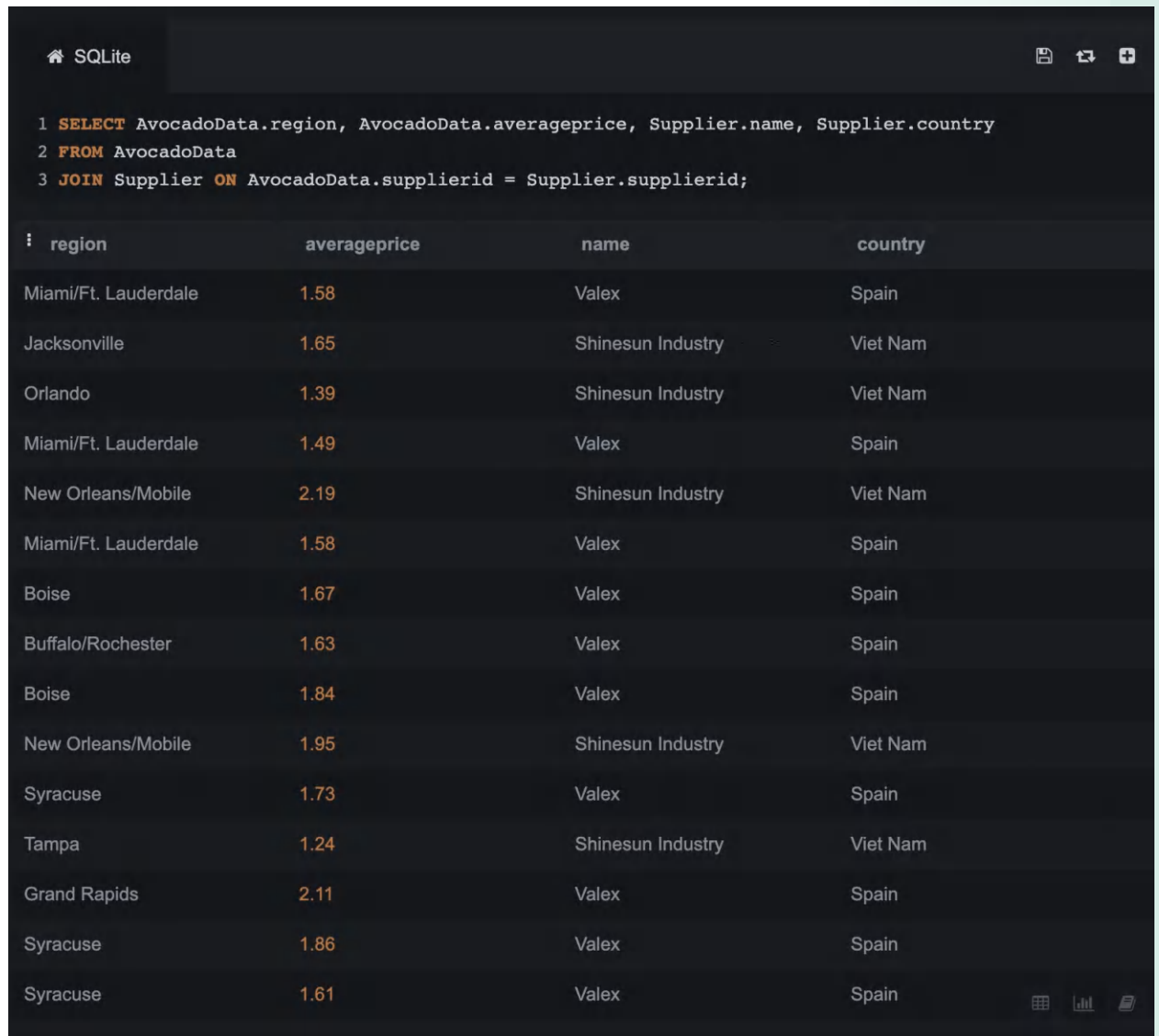
```

SELECT AvocadoData.region, AvocadoData.averageprice, Supplier.name, Supplier.
country
FROM AvocadoData
JOIN Supplier ON AvocadoData.supplierid = Supplier.supplierid;

```

This table provides data in a cohesive form. We can now see that Valex from Spain supplied the product to the Miami/Ft. Lauderdale region where the average customer price was \$1.58.

Exhibit 2-27



The screenshot shows an SQLite application window with a dark theme. At the top, there is a title bar with a home icon, the text 'SQLite', and three utility icons (save, undo, redo). Below the title bar, a SQL query is entered in a text area:

```
1 SELECT AvocadoData.region, AvocadoData.averageprice, Supplier.name, Supplier.country
2 FROM AvocadoData
3 JOIN Supplier ON AvocadoData.supplierid = Supplier.supplierid;
```

Below the query, the results are displayed in a table with four columns: region, averageprice, name, and country. The table contains 15 rows of data. At the bottom right of the window, there are three icons: a grid, a bar chart, and a document.

region	averageprice	name	country
Miami/Ft. Lauderdale	1.58	Valex	Spain
Jacksonville	1.65	Shinesun Industry	Viet Nam
Orlando	1.39	Shinesun Industry	Viet Nam
Miami/Ft. Lauderdale	1.49	Valex	Spain
New Orleans/Mobile	2.19	Shinesun Industry	Viet Nam
Miami/Ft. Lauderdale	1.58	Valex	Spain
Boise	1.67	Valex	Spain
Buffalo/Rochester	1.63	Valex	Spain
Boise	1.84	Valex	Spain
New Orleans/Mobile	1.95	Shinesun Industry	Viet Nam
Syracuse	1.73	Valex	Spain
Tampa	1.24	Shinesun Industry	Viet Nam
Grand Rapids	2.11	Valex	Spain
Syracuse	1.86	Valex	Spain
Syracuse	1.61	Valex	Spain

SQLite

Update the Data

Step 17: Using SQL, you can change the values in any row and columns. For example, Supplier A has started a new operation in Mexico that it will now be using to provide avocados to the U.S. market. In this instance, the company's location will need to be changed from Spain to Mexico. Perform this task by using the following query statement (Exhibit 2-28):

```
UPDATE Supplier SET country = 'Mexico' WHERE name = 'Valex';
```

Review your changes by using the following query statement:

```
Select* FROM Supplier;
```

Exhibit 2-28

SQLite

1 SELECT* FROM Supplier;

supplierid	name	country
A	Valex	Mexico
B	Cresco Produce	Thailand
C	Shinesun Industry	Viet Nam

SQLite

Delete Values

Step 18: You might find that you no longer need to retain a row of data. Therefore, you must delete it to prevent it from being included in future analyses. To illustrate this, let’s start with displaying observationid 5. Perform this task by using the following query statement (Exhibit 2-29):

Select * from AvocadoData WHERE observationid = ‘5’;

You now see observation 5 is visible on the screen.

Exhibit 2-29

SQLite

1 SELECT * FROM AvocadoData WHERE observationid = '5';

observationid	region	year	month	quarter	type	averageprice	totalvolume	supplierid
5	Boston	2019	3	1	Conventional	1.32	835837.52	B

SQLite

Step 19: We want to delete this observation from the data. Perform this task by using the following query statement (Exhibit 2-30):

Delete from AvocadoData WHERE observationid = ‘5’;

Select * from AvocadoData;

As you see, observation 5 no longer exists in the table.

Exhibit 2-30

🏠 SQLite

📄

↺↻

⊕

1

SELECT * FROM AvocadoData;

i	observationid	region	year	month	quarter	type	averageprice	totalvol...	supplierid
	17038	Miami/Ft. La...	2016	10	4	Organic	1.58	385.55	A
	6381	Jacksonville	2018	11	4	Organic	1.65	404.62	C
	6392	Orlando	2018	11	4	Organic	1.39	405.29	C
	16826	Miami/Ft. La...	2016	10	4	Organic	1.49	472.82	A
	10471	New Orleans...	2017	6	2	Organic	2.19	515.01	C
	16985	Miami/Ft. La...	2016	10	4	Organic	1.58	542.85	A
	15109	Boise	2016	2	1	Organic	1.67	562.64	A
	17019	Buffalo/Roch...	2016	10	4	Organic	1.63	563.06	A
	15162	Boise	2016	2	1	Organic	1.84	566.57	A
	10153	New Orleans...	2017	4	2	Organic	1.95	634.09	C
	17063	Syracuse	2016	10	4	Organic	1.73	667.95	A
	6411	Tampa	2018	11	4	Organic	1.24	710.36	C
	17028	Grand Rapids	2016	10	4	Organic	2.11	740.33	A
	17275	Syracuse	2016	11	4	Organic	1.86	775.8	A
	16957	Syracuse	2016	10	4	Organic	1.61	795.95	A
	14943	Syracuse	2016	1	1	Organic	1.64	809.41	A

📊

📈

📄

SQLite

In this section, you have learned how to query data using SELECT in one table and across tables using JOIN, how to create a table using CREATE and INSERT, and how to modify data using UPDATE. Each query is helpful in identifying different pieces of information. By experiencing SQL, you have discovered how databases work. It is important to remember that the ideas you have learned here are just the beginning of your journey with SQL.

Data Is the Currency of Marketing

Data is the currency of marketing. But so is marketing analytics, particularly in the era of big data. Data is only one aspect of answering marketing questions or solving marketing problems. Marketers also need analytics to examine their data and convert it to marketing knowledge. For example, to answer marketing questions like “How do you know that?” and “What can you do with that to improve your marketing decision-making?”, you must have both data and analytical methods skills.

Marketing analytics is of little value if you do not have quality data, and data is of little value without analytics. A decade ago, marketing analysts could make effective use of

data for decision-making with means, medians, standard deviations, correlations, differences tests, and occasionally multiple regression. Today, much more advanced analytical tools are essential to understand data and make decisions that ensure a company's competitiveness in the marketplace. This is true not only for consumer marketing, but also for business-to-business marketing. It is also true across all types of industries, and for nonprofit organizations as well as profit-oriented business ventures.

Consider, for example, what you have learned in this chapter. Data is not just numbers, it must be identified and organized to make it usable. The case study in this chapter is not only about avocados, but what data you might need, how to prepare your data to make it usable for decision-making, and then presenting it so marketers can understand and apply it quickly. As we pointed out in the chapter, the first step is data preparation and cleaning so the data will be usable. This involves dealing with missing data, correcting errors, and so forth. Next is to organize the data into a table format, which you learned how to do in this chapter for relational databases using Structured Query Language (SQL).

To make data usable, marketers apply analytics so they can understand the data and make decisions with it. For the avocado toast case example in this chapter, you learned that data like price, product characteristics (organic vs. non-organic), sales by region (city, state, etc.), and time of year (when available and from what sources/suppliers) are important for vendors selling produce, such as avocados. If the data is not organized, it is not usable, and vendors cannot sell it. Relational databases are the tool to do this.

On the other side of this transaction, purchasers of avocados must have this same information. For example, before including "Avocado Toast" on the menu, restaurant managers must be confident they can obtain a sufficient amount of avocados to serve their customers, from which source, and at what price? Thus, for both the vendor selling the product and the purchaser using the product, organizing and presenting the information to make decisions is essential. Understanding SQL or similar analytical methods to communicate the data is, therefore, a fundamental and essential tool for marketers. The avocado toast example in this chapter is simple. It has less than 10 items to keep track of, including price, when and how much is available, and from whom, organic versus non-organic avocados to meet customer preferences, and so on. But most business transactions involve many more variables, often in the hundreds or sometimes the thousands. The retrieval of data is thus an essential skill and tool for marketing analytics to be useful in decision-making.

We encourage you to find more ways to work with SQL. Here are a few options for you to explore:

- *Google's BigQuery*: This Google database solution does not require a server and allows you to use SQL-like language to query and manipulate the data. BigQuery allows analysis in real time. It allows for 1 TB of data and 10 GB of data for free each month. To learn more, visit the Google BigQuery documentation at <https://cloud.google.com/bigquery>.
- *MYSQL on the server-side*: SQL can be used to manage data that is on the server side. You can set up a server-side database and access it using a client-based MySQL workbench used for data modeling, SQL development, and user administration. To learn more, review the MySQL documentation at <https://www.mysql.com/products/workbench/>
- *Amazon Aurora*: This relational database on Amazon Web Services is five times faster than MYSQL. Aurora is set up using Amazon Relational Database Service (RDS) on the virtual server. Data can be loaded into Aurora from MySQL and PostgreSQL. The data is backed up continuously on Amazon S3 servers to ensure reliability disaster recovery. To learn more, review the Aurora document at <https://aws.amazon.com/rds/aurora/getting-started>.

Summary of Learning Objectives and Key Terms

LEARNING OBJECTIVES

- Objective 2.1** Define big data and summarize the journey from big data to smart data.
- Objective 2.2** Discuss database management systems, relational databases, and SQL query language.
- Objective 2.3** Investigate the key elements of enterprise data architecture.
- Objective 2.4** Define the dimensions of data quality and describe the importance of performing marketing analytics.
- Objective 2.5** Explain the importance of understanding and preparing data prior to engaging in analytics.

KEY TERMS

Aggregation	Database	Smart data
Big data	Feature	Streaming data
Categorical variables	Hadoop	Structured Query Language (SQL)
Cluster analysis	Normalization	Unit of analysis
Data lake	Outlier	Value
Data management	Overfitting	Variety
Data mart	Power calculation	Velocity
Data quality	Relational database	Veracity
Data warehouse	Sample size	Volume

Discussion and Review Questions

1. Define and describe the characteristics of big data.
2. Why is it important for marketers to have a basic understanding of the fundamentals surrounding data management?
3. What are some characteristics of data quality that need to be examined to avoid invalid analytics results?
4. How is the data prepared in ETL?
5. Explain several tasks that might be involved when preparing data for analysis.
6. What are some basic questions you can answer by querying data from a relational database using SQL?

Critical Thinking and Marketing Applications

1. Congratulations, you were just hired as a marketing analyst for a large company. The VP of Marketing has asked you to examine how the company might improve sales. What data might be helpful in your exploration? Where might you locate the data needed? What questions should you ask first?

2. Consider the avocado data within the Case Study. What additional data fields might be necessary to explore the following questions:
 - a. What is the average income of customers in the cities that yield the largest sales?
 - b. Do weather patterns impact the sale of avocados each week?
 - c. Does social media chatter influence the purchase of weekly avocado sales?

References

1. Bernard Marr, "The Amazing Ways Coca-Cola Uses Artificial Intelligence and Big Data to Drive Success," *Forbes*, September 18, 2017, <https://www.forbes.com/sites/bernardmarr/2017/09/18/the-amazing-ways-coca-cola-uses-artificial-intelligence-ai-and-big-data-to-drive-success> (accessed June 15, 2019); and Bernard Marr, "How Coca-Cola Is Using AI to Stay at the Top of the Soft Drinks Market," *AI News*, May 7, 2019, <https://www.artificialintelligence-news.com/2019/05/07/how-coca-cola-is-using-ai-to-stay-at-the-top-of-the-soft-drinks-market> (accessed June 15, 2019).
2. Bernard Marr, "The Amazing Ways eBay Is Using Artificial Intelligence to Boost Business Success," *Forbes*, April 26, 2019, <https://www.forbes.com/sites/bernardmarr/2019/04/26/the-amazing-ways-ebay-is-using-artificial-intelligence-to-boost-business-success> (accessed June 15, 2019); and Sanjeev Katariya, "eBay's Platform Is Powered by AI and Fueled by Customer Input," *eBay*, March 13, 2019, <https://www.ebayinc.com/stories/news/ebays-platform-is-powered-by-ai-and-fueled-by-customer-input> (accessed June 15, 2019).
3. Bernard Marr, "The Amazing Ways How Mastercard Uses Artificial Intelligence to Stop Fraud and Reduce False Declines," *Forbes*, November 30, 2018, <https://www.forbes.com/sites/bernardmarr/2018/11/30/the-amazing-ways-how-mastercard-uses-artificial-intelligence-to-stop-fraud-and-reduce-false-declines> (accessed June 15, 2019); and Clinton Boulton, "3 Ways Mastercard Uses AI to Fight Fraud," *CIO*, December 3, 2018, <https://www.cio.com/article/3322927/3-ways-mastercard-uses-ai-to-fight-fraud.html> (accessed June 15, 2019).
4. Bob Violino, "The Secrets of Highly Successful Data Analytics Teams," *Insider Pro*, October 24, 2017, <https://www.idginsiderpro.com/article/3234353/the-secrets-of-highly-successful-data-analytics-teams.html> (accessed June 15, 2019); and Phil Weinzimer, "How CIO-CMO Partnerships Leverage Omni-Channel Marketing Strategy to Drive Business Value," *CIO*, February 21, 2017, <https://www.cio.com/article/3171075/how-cio-cmo-partnerships-leverage-omni-channel-marketing-strategy-to-drive-business-value.html> (accessed June 15, 2019).
5. Danny Bradbury, "How Big Data in Aviation Is Transforming the Industry," *Cloudera*, <https://hortonworks.com/article/how-big-data-in-aviation-is-transforming-the-industry> (accessed June 15, 2019); and Oliver Wyman, "The Data Science Revolution That's Transforming Aviation," *Forbes*, June 16, 2017, <https://www.forbes.com/sites/oliverwyman/2017/06/16/the-data-science-revolution-transforming-aviation> (accessed June 15, 2019).
6. J. Clement, "Combined Desktop and Mobile Visits to Target.com from May 2019 to February 2020," *Statista*, March 19, 2020, <https://www.statista.com/statistics/714572/web-visits-to-targetcom> (accessed June 15, 2019).
7. "Examples of Data Volumes," University of Delaware, <https://www.eecis.udel.edu/~amer/Table-Kilo-Mega-Giga-YottaBytes.html> (accessed June 15, 2019).
8. J.D. Byrum, "The Grocery List: Why 140 Million Americans Choose Walmart," *Walmart*, October 3, 2016, <https://blog.walmart.com/business/20161003/the-grocery-list-why-140-million-americans-choose-walmart> (accessed June 15, 2019); and Krishna Thakker, "Kroger and Walmart Outline Digital Transformations," *Grocery Dive*, January 14, 2019, <https://www.grocerydive.com/news/kroger-and-walmart-outline-digital-transformations/545947> (accessed June 15, 2019).
9. Clint Boulton, "5 Data Analytics Success Stories: An Inside Look," *CIO*, February 25, 2020, <https://www.cio.com/article/3221621/6-data-analytics-success-stories-an-inside-look.html> (accessed June 15, 2019); Doug Henschen, "Merck Optimizes Manufacturing With Big

InformationWeek, April 2, 2014, <https://www.informationweek.com/strategic-cio/executive-insights-and-innovation/merck-optimizes-manufacturing-with-big-data-analytics/d/d-id/1127901> (accessed June 15, 2019); and Ken Murphy, “Merck Focuses on Data-Driven Platform Strategy,” *SAPinsider*, <https://sapinsider.wispubs.com/Assets/Case-Studies/2016/December/IP-Merck-Focuses-on-Data-Driven-Platform-Strategy> (accessed June 15, 2019).

10. Penny Crossman, “TD Bank’s Bold Bet on AI,” *American Banker*, January 16, 2018, <https://www.americanbanker.com/news/td-bank-investments-builds-on-ai-strategy> (accessed June 15, 2019); Bernard Marr, “The Amazing Ways TD Bank, Canada’s Second-Largest Bank, Uses Big Data, AI & Machine Learning,” *Forbes*, December 18, 2018, <https://www.forbes.com/sites/bernardmarr/2018/12/18/the-amazing-ways-td-bank-canadas-second-largest-bank-uses-big-data-ai-machine-learning> (accessed June 15, 2019).
11. Joseph F. Hair, William C. Black, Barry J. Babin, and Rolph E. Anderson, *Multivariate Data Analysis*, 8th ed. (EMEA: Cengage Learning, 2019).
12. Hass Avocado Board, <https://hassavocadoboard.com> (accessed June 15, 2019).